Università degli Studi di Catania
Dipartimento di Matematica ed Informatica

Dottorato di Ricerca in Informatica

XXXIII Ciclo

Multimedia Forensics: From Image manipulation to the Deep Fake. New Threats in the Social Media Era.

Cristina Nastasi

Advisor
Prof. Sebastiano Battiato

Università degli Studi di Catania
Dipartimento di Matematica ed Informatica

Dottorato di Ricerca in Informatica

XXXIII Ciclo

Settore scientifico-disciplinare INF/01

Multimedia Forensic: From Image manipulation to the Deep Fake. New Threats in the Social Media Era

Cristina Nastasi

Advisor
Prof. Sebastiano Battiato

Author's address:
Cristina Nastasi
Dipartimento di Matematica ed Informatica
Università degli Studi di Catania
V.le A. Doria, 6 I-95125 Catania, Italy
e-mail: `cristina.nastasi@unict.it`
www: `http://www.unict.it`

Dedicated to my Family

Introduction

This dissertation reports the research activities carried on during the Ph.D. program in Computer Science I attended at the University of Catania – Departments of Mathematics and Computer Science. My research area is mainly concerned with Social and Multimedia Forensics, starting from a survey and a forensic analysis conduct on 250 Social Network, passing through a detection of false multimedia content and manipulated images (focusing particularly on the double compressed JPEG images), until arriving to the emerging phenomenon of the DeepFakes. Nowadays the social media ecosystem includes hundreds of applications, based on web and mobile technologies and able to allow people to communicate easily and/or to share information, resources, images and videos. However it also hides a serious risk of disclosure of personal data that can lead to privacy issues and crimes. So it is very important to study social network applications also from a forensic point of view. This work aims to explore the Social Media ecosystems [1], formed by 250 web and mobile applications, analyzing everyone to exploit and provide a complete survey on the forensic methods that can be applied to recover useful information (ID user, ID post, . . .) to be used in front of a court. In the second part of our research, we focus the attention on the multimedia content and particularly on JPEG images. One of the most common problem in the image forensic field is the reconstruction of the history of an image. Image Forensics analysis is the process that aims to understand the history of a digital image. In the last ten years Image Forensics, whose goal is exploit the knowledge from the science of Image Processing to answer questions that arise in the forensic scenario, has developed at a growing rhythm in order to efficiently check the originality of images and videos from the all-day context. The reconstruction data of the further processing are very important information

that can help us to have some useful hints about the originality of the image under analysis. If an image has been subjected to more than one JPEG compression the considered image is not the exact bitstream generated by the camera at the time of shooting. The originality of an image can be established by recovering coefficients of the JPEG quantization matrix used to compress an image at the time of shooting (i.e., when the image has been created), when, for some reasons, this information is no more available in the Exif metadata. This scenario may include the primary quantization coefficients of an image that has been doubly JPEG compressed, or the retrieval of the compression matrix of an uncompressed image previously JPEG compressed, since in both these cases the values of the primary compression steps are lost. So it is clear that Double Quantization Detection (DQD) and Quantization Step Estimation (QSE) are two fundamental steps in the overall process. This research deals with also the Double JPEG compression detection algorithm. Starting from the paper of Galvan et al. [59] in this work is presented a Matlab implementation of the algorithm by adopting a different approach based on a block splitting methodologies and the results of this algorithm applied on different contexts: high resolution images, minor block number... This PhD thesis investigate finally one of the most terrifying emerging phenomenon in the digital world: the Deepfake. It presents a brief overview of the technologies able to automatically replace a person's face in images and videos by exploiting algorithm based on deep learning and also several techniques of creation and detection of the so-called Deepfakes with the related social and legal problems. A forensic analysis of those images with standard method will be presented and will show that not surprisingly state of the art techniques are not completely able to detect the fakeness. It is very important to be able to counter this phenomenon by creating new forensic methods. Moreover, this work shows an idea on how to fight Deepfake images by analyzing anomalies in the frequency domain and evaluating details and traces of

underlying generation process of the image (e.g. in the Fourier domain).

This Phd thesis provides a good study of the state of the art in Social and Multimedia Forensics and shows some results obtained through an in-depth analysis.

Contents

Chapter 1

From Digital Evidence to Digital Forensic

1.1 The Notion of Digital Evidence

Nowadays digital devices are everywhere in our life by helping people to communicate locally and globally easily. When you speak about a digital crime most people immediately think of computers as the only sources for digital evidence. But computers are not the only sources of evidence, any piece of technology that processes information can be used in a criminal way (cell phones, Internet, . . .) and therefore there are many sources of digital evidence. Digital forensic evidence can come from any electronic storage or communications media such as cellphones, computers, iPod's, video game consoles etc, but it is fragile in fact it can be easily damaged or altered due to improper handling, whether by accident or on purpose. The different types of crimes tend to lend themselves to one device or the other.

A key component of the digital investigative process involves the assessment of potential evidence in a cyber crime. Central to the effective processing of evidence is a clear understanding of the details of the case at hand and thus, the classification of cyber crime in question. For instance, if an agency seeks to prove that an individual has committed crimes related to identity theft, digital forensics investigators use sophisticated methods to analyse hard drives, email accounts, social networking sites, and other digital archives to retrieve and assess any information that can serve as viable evidence of the crime. Prior to conducting an investigation, the investigator must define the types of evidence to analyse (including specific platforms and data formats) and have a clear understanding of how to preserve pertinent data. The investigator must then determine the source and integrity of such data before entering it into evidence.

The important thing to know is that need to be able to recognize and properly seize potential digital evidence. Digital evidence is defined as information and data of value to an investigation that is stored on, received or transmitted by an electronic device [2]. This evidence can be acquired when electronic devices are seized and secured for examination. Digital evidence is latent (hidden), like fingerprints or DNA evidence, crosses jurisdictional borders quickly and easily, can be altered, damaged or destroyed with little effort and can be time sensitive.

The area of informatics is constantly evolving and therefore it is difficult to introduce long-term stable and applicable standards to this area. Procedures and recommendations for collecting digital traces are evolving not only along the increasing number of investigations, but also along with the development of digital devices. Procedures of a similar nature should ideally be frequently updated, in order to allow to factor any new circumstances that may occur during the investigation.

Analysis and interpretation of Digital Evidence can be a complex process. Individual investigators, organizations and jurisdictions may use different approaches, methods, processes and controls to conduct a digital investigation.

One of the most critical issues in forensic investigations is the acquisition and preservation of evidence in such a way as to ensure its integrity. As with conventional physical evidence, it is crucial for the first and subsequent responders (defined as "Digital Evidence First Responders" and "Digital Evidence Specialists") to maintain the chain of custody of all digital forensic evidence, ensuring that it is gathered and protected through structured processes that are acceptable to the courts. More than simply providing integrity, the processes must provide assurance that nothing untoward can have occurred. This requires that a defined baseline level of information security controls is met or exceeded.

It is hoped that standardization will lead to the adoption of similar, if not identical, approaches internationally, making it easier to compare, combine and contrast the results of such investigations even when performed by different people or organizations and potentially across different jurisdictions.

A survey of the International Digital Forensic Standard wil be considered in the next section.

1.2 The International Standardization Approach in Digital Forensic Domain

ISO (the International Organization for Standardization) and IEC (the lnternational Electrotechnical Commission) represent the specialized system for worldwide standardization and provide guidelines for specific activities, also in the field of the Digital Forensic. National bodies that are members of lSO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1. Promote good practice methods and processes for forensic capture and investigation of digital evidence is the fundamental purpose of the ISO 27k digital forensics standards family.

A guidelines for specific activities in handling digital evidence, which are identification, collection, acquisition and preservation of digital evidence that may be of evidential value, can be found on the International Standard ***ISO/IEC 27037*** [3].

Prior to the release of ISO/IEC 27037, there were no globally accepted standards on acquiring digital evidence, the first step in the process. Police have developed their own national guidelines and procedures for the acquisition and protection of electronic evidence. However, this creates issues when cross-border crimes are committed since digital forensic evidence acquired in one country may need to be presented in the courts of another. The evidence that may have been acquired or preserved without the right level of security may be legally inadmissible. This standard provides detailed guidance on the identification, collection and/or acquisition, marking, storage, transport and preservation of electronic evidence, particularly to maintain its integrity. It defines and describes the processes through which evidence is recognized and identified, documentation of the crime scene, collection and preservation of the evidence, and the packaging and transportation of evidence. These processes are required in an investigation that is designed to maintain the integrity of the digital evidence – an acceptable methodology in obtaining digital evidence that will contribute to its admissibility in legal and disciplinary actions as well as other required instances. This International Standard also provides general guidelines for the collection of non-digital evidence that may be helpful in the analysis stage of the potential digital evidence.

The scope covers "traditional" IT systems and media rather than vehicle systems, cloud computing etc. The guidance is aimed primarily at first responders. It intends to provide guidance to those individuals responsible for the identification, collection, acquisition and preservation of potential digital evidence. These individuals include Digital Evidence First Responders (DEFRs), Digital Evidence Specialists (DESs), incident response specialists and forensic laboratory managers. This International Standard ensures that responsible individuals manage potential digital evidence in practical ways that are acceptable worldwide, with the objective to facilitate investigation involving digital devices and digital evidence in a systematic and impartial manner while preserving its integrity and authenticity.

Every country has its own unique legislative system. A crime committed in one jurisdiction may not even be regarded as a crime in another. The challenge is to harmonize processes across borders such that cybercriminals can be prosecuted accordingly. Therefore, a means to allow and facilitate the exchange and use of reliable evidence (i.e. an international standard on acquiring digital evidence) is required. "Digital evidence", meaning information from digital devices to be presented in court, is interpreted differently in different jurisdictions. For the widest applicability, the standard will avoid using jurisdiction-specific terminology. This International Standard provides guidance to individuals with respect to common situations encountered throughout the digital evidence handling process and assists organizations in their disciplinary procedures and in facilitating the exchange of potential digital evidence between jurisdictions.

This International Standard also intends to inform decision-makers who need to determine the reliability of digital evidence presented to them. It is applicable to organizations needing to protect, analyze and present potential digital evidence. It is relevant to policy-making bodies that create and evaluate procedures relating to digital evidence, often as part of a larger body of evidence.

Like any standard of ”best practices”, also ISO 27037 is a living standard and its practical application is always an adaptation of recommendations to specific conditions, so that conflicts between the demands of standard and practice is avoided. It is also clear that the dynamics of the development of information technology is greater than the ability to update a particular ISO standard.

The standard should also serve as a basis for assessing qualifications needed for expert practice. This is because actual practice shows that many judicial experts are not acquainted even with the basic requirements for collection and acquisition of digital evidence, which are listed at the beginning of the standards. Judicial experts can find themselves in different positions – for example as consultants in police work, as independent experts, or they can work as experts in their (own) laboratories. The standard should therefore be one of the most important documents when considering qualification requirements of ICT security specialists. Since its creation, the standard ISO 27037 has had some – perhaps cardinal – shortcomings. One can wonder why it has not also been based on other documents and recommendations that had long been accepted by the international community of digital forensics experts (e.g.

recommendations of Association of Chief Police Officer, Digital Forensic Research Workshop, and International Organization for Cooperation in Evaluation, European Network of Forensic Science Institutes). Nevertheless, the standard unified at least the rudimentary principles and rules of providing digital evidence. Many of those procedures seem obvious and such that would already be used in practice. This standard could also give rise to regulatory measures, because providing digital evidence is a process from which ensues obtaining of evidence. This fact is so significant that some elementary rules should never be violated. If a biological sample is proved to be contaminated, it cannot be used as evidence. Paradoxically, the violation of fundamental procedures in working with digital evidence does not exclude the contaminated data from evidentiary proceedings. In conclusion, it can be stated that the standard is understood as a convenient first step towards a process that will lead to increased awareness of the specifics of working with digital evidence – despite some of its obvious shortcomings. This standard also contributes to a gradual improvement in the work of all people who come into contact with digital evidence, whether they be police technicians, legal experts, investigators, lawyers, judges, as well as security specialists and ICT administrators or members of CERT.

This International Standard complements ISO/IEC 27001 and ISO/IEC 27002, and in particular the control requirements concerning potential digital evidence acquisition by providing additional implementation guidance. In addition, this International Standard will have applications in contexts independent of ISO/IEC 27001 and ISO/IEC 27002. This International Standard should be read in conjunction with other standards related to digital evidence and the investigation of information security incidents.

This International Standard gives guidance for the following devices (the below list presented is an indicative list and not exhaustive) and/or functions that are used in various circumstances:

- Digital storage media used in standard computers like hard drives, floppy disks, optical and magneto optical disks, data devices with similar functions,
- Mobile phones, Personal Digital Assistants (PDAs), Personal Electronic Devices (PEDs), memory cards,
- Mobile navigation systems,
- Digital still and video cameras (including CCTV),
- Standard computer with network connections,
- Networks based on TCP/IP and other digital protocols, and
- Devices with similar functions as above.

Due to the fragility of digital evidence, it is necessary to carry out an acceptable methodology to ensure the integrity and authenticity of the potential digital evidence. This International Standard does not mandate the use of particular tools or methods. Key components that provide credibility in the investigation are the methodology applied during the process, and individuals qualified in performing the tasks specified in the methodology.

The ISO 27037 will not cover analysis of digital evidence, nor its admissibility, weight, relevance etc. This International Standard does not address the methodology for legal proceedings, disciplinary procedures and other related actions in handling potential digital evidence that are outside the scope of identification, collection, acquisition and preservation.

The lnternational Standard ***ISO/IEC 27042*** [4] provides guidance on the conduct of the analysis and interpretation of potential digital evidence in order to identify and evaluate digital evidence which can be used to aid understanding of an incident. The exact nature of the data and information making up the potential digital evidence will depend on the nature of the incident and the digital evidence sources involved in that incident. This International Standard assumes that the guidance given in ISO/IEC 27035-2 and ISO/IEC 27037 has been followed and that all processes used are compatible with the guidance given in ISO/IEC 27043 and ISO/IEC 27041.

It provides guidance on the analysis and interpretation of digital evidence in a manner which addresses issues of continuity, validity, reproducibility, and repeatability. It encapsulates best practice for selection, design, and implementation of analytical processes and recording sufficient information to allow such processes to be subjected to independent scrutiny when required. It provides guidance on appropriate mechanisms for demonstrating proficiency and competence of the investigative team. In some circumstances, there can be several methods which could be applied and members of the investigative team will be required to justify their selection of a particular process and show how it is equivalent to another process used by other investigators. In other circumstances, investigators may have to devise new methods for examining digitai evidence which has not previously been considered and should be able to show that the method produced is ”fit far purpose”. Application of a particular method can influence the interpretation of digital evidence processed by that method. The available digital evidence can influence the selection of methods far further analysis of digital evidence which has already been acquired. The ISO/IEC 27042 provides a common framework, far the analytical and interpretational elements of information systems security incident handling, which can be used to assist in the implementation of new methods and provide a minimum common standard far digital evidence produced from such activities. This International Standard is intended to complement other standards and documents which give guidance on the investigation of, and preparation to investigate, information security incidents. It is not a comprehensive guide, but lays down certain fundamental principles which are intended to ensure that tools, techniques, and methods can be selected appropriately and shown to be fit for purpose should the need arise. This International Standard also intends to inform decision-makers that need to determine the reliability of digital evidence presented to them. It is applicable to organizations needing to protect, analyse, and present potential digital evidence. It is relevant to policy-making bodies that create and evaluate procedures relating to digital evidence, often as part of a larger body of evidence. This International Standard describes part of a comprehensive investigative process which includes, but is not limited to, the following topic areas:

- incident management, including preparation, and planning for investigations;
- handling of digital evidence;
- use of, and issues caused by, redaction;
- intrusion prevention and detection systems, including information which can be obtained from these systems;
- security of storage, including sanitization of storage;
- ensuring that investigative methods are fit for purpose;
- carrying out analysis and interpretation of digital evidence;
- understanding principles and processes of digital evidence investigations;
- security incident event management, including derivation of evidence from systems involved in security incident event management;
- relationship between electronic discovery and other investigative methods, as well as the use of electronic discovery techniques in other investigations;

- governance of investigations, including forensic investigations.

These topic areas are addressed, in part, by the following ISO/IEC standards.

ISO/IEC 27037

This International Standard (described in depth above) describes the means by which those involved in the early stages of an investigation, including initial response, can assure that sufficient potential digital evidence is captured to allow the investigation to proceed appropriately.

ISO/IEC 27038

Some documents can contain information that must not be disclosed to some communities. Modified documents can be released to these communities after an appropriate processing of the original document. The process of removing information that is not to be disclosed is called “redaction”. The digital redaction of documents is a relatively new area of document management practice, raising unique issues and potential risks. Where digital documents are redacted, removed information must not be recoverable. Hence, care needs to be taken so that redacted information is permanently removed from the digital document (e.g. it must not be simply hidden within non-displayable portions of the document). ISO/IEC 27038 specifies methods for digital redaction of digital documents. It also specifies requirements for software that can be used for redaction.

ISO/IEC 27040:2015

This International Standard provides detailed technical guidance on how organizations can define an appropriate level of risk mitigation by employing a well-proven and consistent approach to the planning, design, documentation, and implementation of data storage security. Storage security applies to the protection (security) of information where it is stored and to the security of the information being transferred across the communication links associated with storage. Storage security includes the security of devices and media, the security of management activities related to the devices and media, the security of applications and services, and security relevant to end-users during the lifetime of devices and media and after end of use. Security mechanisms like encryption and sanitization can affect one's ability to investigate by introducing obfuscation mechanisms. They have to be considered prior to and during the conduct of an investigation. They can also be important in ensuring that storage of evidential material during and after an investigation is adequately prepared and secured.

ISO/IEC 27041

It is important that methods and processes deployed during an investigation can be shown to be appropriate. This International Standard provides guidance on how to provide assurance that methods and processes meet the requirements of the investigation and have been appropriately tested.

ISO/IEC 27043:2015

This International Standard defines the key common principles and processes underlying the investigation of incidents and provides a framework model for all stages of investigations. The following ISO/IEC projects also address, in part, the topic areas identified above and can lead to the publication of relevant standards at some time after the publications of this International Standard.

ISO/IEC 27035 (all parts)

This is a three-part standard that provides organizations with a structured and planned approach to the management of security incident management. It is composed of

ISO/IEC 27035-1

This part presents basic concepts and phases of information security incident management. It combines these concepts with principles in a structured approach to detecting, reporting, assessing, responding, and applying lessons learned.

ISO/IEC 27035-2

This part presents the concepts to plan and prepare for incident response. The concepts, including incident management policy and plan, incident response team establishment, and awareness briefing and training, are based on the plan and prepare phase of the model presented in ISO/IEC 27035-1. This part also covers the "Lessons Learned" phase of the model.

ISO/IEC 27035-3

This part includes staff responsibilities and practical incident response activities across the organization. Particular focus is given to the incident response team activities such including monitoring, detection, analysis, and response activities for the collected data or security events.

ISO/IEC 27044

This provides guidelines to organizations in preparing to deploy security information and event management processes/systems. In particular, it addresses the selection, deployment, and operations of SIEM. It intends specifically to offer assistance in satisfying requirements of ISO/IEC 27001 regarding the implementation of procedures and other controls capable of enabling prompt detection and response to security incidents, to execute monitoring, and review procedures to properly identify attempted and successful security breaches and incidents.

ISO/IEC 27050 (all parts)

This addresses activities in electronic discovery, including, but not limited to identification, preservation, collection, processing, review, analysis, and production of electronically stored information (ESI). In addition, it provides guidance on measures, spanning from initial creation of ESI through its final disposition, which an organization can undertake to mitigate risk and expense should electronic discovery become an issue. It is relevant to both non-technical and technical personnel involved in some or all of the electronic discovery activities. It is important to note that this guidance is not intended to contradict or supersede local jurisdictional laws and regulations. Electronic discovery often serves as a driver for investigations, as well as evidence acquisition and handling activities. In addition, the sensitivity and criticality of the data sometimes necessitate protections like storage security to guard against data breaches.

ISO/IEC 30121:2015

This International Standard provides a framework for governing bodies of organizations (including owners, board members, directors, partners, senior executives, or similar) on the best way to prepare an organization for digital investigations before they occur. This International Standard applies to the development of strategic processes (and decisions) relating to the retention, availability, access, and cost effectiveness of digital evidence disclosure. This International Standard is applicable to all types and sizes of organizations. The International Standard is about the prudent strategic preparation for digital investigation of an organization. Forensic readiness assures that an organization has made the appropriate and relevant strategic preparation for accepting potential events of an evidential nature. Actions can occur as the result of inevitable security breaches, fraud, and reputation assertion. In every situation, information technology (IT) has to be strategically deployed to maximize the effectiveness of evidential availability, accessibility, and cost efficiency.

1.3 Steps of Digital Forensic

The field of digital forensics investigation is growing, especially as law enforcement and legal entities realize just how valuable information technology (IT) professionals are when it comes to investigative procedures. With the advent of cyber crime, tracking malicious online activity has become crucial for protecting private citizens, as well as preserving online operations in public safety, national security, government and law enforcement. Tracking digital activity allows investigators to connect cyber communications and digitally-stored information to physical evidence of criminal activity; computer forensics also allows investigators to uncover premeditated criminal intent and may aid in the prevention of future cyber crimes. For those working in the field, there are five basic steps in digital forensics process, all of which contribute to a thorough and revealing investigation.

1. ***Identification*** – The first stage identifies potential sources of relevant evidence/information (devices) as well as key custodians and location of data.

2. ***Preservation and Collection*** – The process of preserving relevant electronically stored information (ESI) by protecting the crime or incident scene, capturing visual images of the scene and documenting all relevant information about the evidence and how collecting digital information that may be relevant to the investigation. Collection may involve removing the electronic device(s) from the crime or incident scene and then imaging, copying or printing out its (their) content.

3. ***Acquisition*** – This is one of the most critical step of successful digital forensic investigation. It is a rigorous, detailed plan for acquiring evidence and consists mainly in producing a forensic copy of the evidence, for preserving digital evidence and protect its authenticity and integrity. The next step, the forensic analysis, are performed on

these digital forensic copies.

4. ***Analysis*** – An in-depth systematic search of evidence relating to the incident being investigated. The outputs of examination are data objects found in the collected information; they may include system- and user-generated files. Analysis aims to draw conclusions based on the evidence found.

5. ***Documentation and Presentation*** – Firstly, reports are based on proven techniques and methodology and secondly, other competent forensic examiners should be able to duplicate and reproduce the same results.

A crucial activity that accompanies the first four steps is contemporaneous note-taking. This is the documentation of what you have done immediately after you have done it in sufficient detail for another person to reproduce what you have done from the notes alone. In the following we explain in depth every single steps .

1.3.1 Identification

The identification process includes searching, detecting and documenting digital evidence (digital evidence is represented in both the physical and the logical form). All devices which could contain digital evidence should be identified in the course of this process. DEFR should carry out a systematic search of the crime scene to prevent overlooking small, camouflaged devices or material which seems irrelevant at first sight. In addition, DEFR should consider the possibility of existence of hidden evidence in the form of virtual components – e.g. Cloud Computing [23]. In case of instability of certain devices (their state may be subject to change over time), it is necessary to prioritize these devices when obtaining evidence. The appropriate order of collection and subsequent processing should minimize possible damage to digital evidence.

1.3.2 Preservation and Collection

In terms of preserving digital evidence, it is necessary to protect its integrity to make it usable for the purposes of investigation. The Digital Evidence First Responders (DEFR) should be able to demonstrate that the evidence has not been changed since its collection and to provide documentation and justification of all actions that led to its changes. After the identification of devices that may contain digital evidence, these devices are removed from their original location and transferred to the laboratory where they are analysed and processed in the next steps. The collection process is at all times documented, including packaging and transport to the laboratory. It is important that DEFR also secures any other physical material which may be related to the digital data that has been collected (e.g. notes on paper which may contain passwords, cradles and power connectors for the devices and other hardware necessary for obtaining digital information from the identified devices).

Collected devices are completely packed (and sealed), preferably in an opaque container (often plastic) so that nobody may manipulate with them. However, the standard ISO/IEC 27037 also recommends (Section 6.9.3.2, 7.1.2.1.2, 7.1.2.1 .3, 7.1.2.2.2, 7.1.2.2.3) placing tape over the power switch and placing tape over the floppy (CD/DVD) disk slot, if present. The Standard divides investigation on procedures for devices that are not connected to the network, and procedures for devices that are connected to the network. This division is no longer necessary, because nowadays almost all the devices are connected to network thus it is required to use only procedures for devices that are connected to the network.

1.3.3 Acquisition

One of the most critical step of successful digital forensic investigation is a rigorous, detailed plan for acquiring evidence. It consists mainly in producing a copy of the evidence (e.g. the content of an entire hard drive) and if necessary, allocated and as well as unallocated space (including deleted files) should be obtained. A very important phase of this step is documenting the methods used. Extensive documentation is needed prior to, during, and after the acquisition process; detailed information must be recorded and preserved, including all hardware and software specifications, any systems used in the investigation process, and the systems being investigated. This step is where policies related to preserving the integrity of potential evidence are most applicable. General guidelines for preserving evidence include the physical removal of storage devices, using controlled boot discs to retrieve sensitive data and ensure functionality, and taking appropriate steps to copy and transfer evidence to the investigator's system. Acquiring evidence must be accomplished in a manner both deliberate and legal. Being able to document and authenticate the chain of evidence is crucial when pursuing a court case, and this is especially true for digital forensics given the complexity of most cases.

Generally, the original and any copies thereof should generate the same output (hash) of the same verification function (proven accurate at that point). Depending on the circumstances (situation, time, price), DEFR should select the most appropriate procedure and method for acquiring data. If this process results in inevitable changes in the created copy, as compared to original, it is necessary to document what data was changed. In those cases in which verification process cannot be carried out (e.g. while acquiring data from a running system, when the original contains bad sectors, or when the time for acquiring data is limited), DEFR should use the best possible way and then be able to justify and vindicate his or her choice of methods. If the created digital copy cannot be verified, then this must be documented and justified. In the case that the source of digital evidence (data) is too big to handle, DEFR can acquire only the relevant part (selected files, folders or paths). All other steps of forensic analysis are performed on copies of digital evidence [77].

The standard in several places mentions the need to create a checksum to verify data, but there is not much stressed that the creation of hash should be recorded in the protocol of digital evidence collection. Such procedure ensures verification that the collected evidence was not tampered.

1.3.4 Analysis

In order to effectively investigate potential evidence, procedures must be in place for retrieving, copying, and storing evidence within appropriate databases. Investigators typically examine data from designated archives, using a variety of methods and approaches to analyze information; these could include utilizing analysis software to search massive archives of data for specific keywords or file types, as well as procedures for retrieving files that have been recently deleted. Data tagged with times and dates is particularly useful to investigators, as are suspicious files or programs that have been encrypted or intentionally hidden. Analyzing file names is also useful, as it can help determine when and where specific data was created, downloaded, or uploaded and can help investigators connect files on storage devices to online data transfers (such as cloud-based storage, email, or other Internet communications). This can also work in reverse order, as file names usually indicate the directory that houses them. Files located online or on other systems often point to the specific server and computer from which they were uploaded, providing investigators with clues as to where the system is located; matching online filenames to a directory on a suspect's hard drive is one way of verifying digital evidence. At this stage, digital forensic investigators work in close collaboration with criminal investigators, lawyers, and other qualified personnel to ensure a thorough understanding of the nuances of the case, permissible investigative actions, and what types of information can serve as evidence.

1.3.5 Documentation and Presentation

In addition to fully documenting information related to hardware and software specs, computer forensic investigators must keep an accurate record of all activity related to the investigation, including all methods used for testing system functionality and retrieving, copying, and storing data, as well as all actions taken to acquire, examine and assess evidence. Not only does this demonstrate how the integrity of user data has been preserved, but it also ensures proper policies and procedures have been adhered to by all parties. As the purpose of the entire process is to acquire data that can be presented as evidence in a court of law, an investigator's failure to accurately document his or her process could compromise the validity of that evidence and ultimately, the case itself. For computer forensic investigators, all actions related to a particular case should be accounted for in a digital format and saved in properly designated archives. This helps ensure the authenticity of any findings by allowing these cybersecurity experts to show exactly when, where, and how evidence was recovered. It also allows experts to confirm the validity of evidence by matching the investigator's digitally recorded documentation to dates and times when this data was accessed by potential suspects via external sources. Now more than ever, cybersecurity experts in this critical role are helping government and law enforcement agencies, corporations and private entities improve their ability to investigate various types of online criminal activity and face a growing array of cyber threats head-on. IT professionals who lead computer forensic investigations are tasked with determining specific cybersecurity needs and effectively allocating resources to address cyber threats and pursue perpetrators of said same. A master's degree in cybersecurity has numerous practical applications that can endow IT professionals with a strong grasp of computer forensics and practices for upholding the chain of custody while documenting digital evidence. Individuals with the talent and education to successfully manage computer forensic investigations may find themselves in a highly advantageous position within a dynamic career field.

1.4 The Visual Evidence and Their Usability in a Court

Digital evidence is information stored or transmitted in binary form that may be relied on in court. It can be found on a computer hard drive, a mobile phone, among other places.

Whether related to malicious cyber activity, criminal conspiracy or the intent to commit a crime, digital evidence can be delicate and highly sensitive. Digital evidence is commonly associated with electronic crime, or e-crime, such as child pornography or credit card fraud. However, digital evidence is now used to prosecute all types of crimes, not just e-crime. For example, suspects' e-mail or mobile phone files might contain critical evidence regarding their intent, their whereabouts at the time of a crime and their relationship with other suspects Cybersecurity professionals understand the value of this information and respect the fact that it can be easily compromised if not properly handled and protected. For this reason, it is critical to establish and follow strict guidelines and procedures for activities related to computer forensic investigations. Such procedures can include detailed instructions about when computer forensics investigators are authorized to recover potential digital evidence, how to properly prepare systems for evidence retrieval, where to store any retrieved evidence, and how to document these activities to help ensure the authenticity of the data. Law enforcement agencies are becoming increasingly reliant on designated IT departments, which are staffed by seasoned digital examiner experts who determine proper investigative protocols and develop rigorous training programs to ensure best practices are followed in a responsible manner. In addition to establishing strict procedures for forensic processes, the digital forensic divisions must also set forth rules of governance for all other digital activity within an organization. This is essential to protecting the data infrastructure of law enforcement agencies as well as other organizations. An integral part of the investigative policies and procedures for law enforcement organizations that utilize computer forensic departments is the codification of a set of explicitly-stated actions regarding what constitutes evidence, where to look for said evidence and how to handle it once it has been retrieved. Prior to any digital investigation, proper steps must be taken to determine the details of the case at hand, as well as to understand all permissible investigative actions in relation to the case; this involves reading case briefs, understanding warrants, and authorizations and obtaining any permissions needed prior to pursuing the case.

Chapter 2

Social Media Forensic

2.1 Introduction on Social Media

Every day million of people connect to the Internet and most of them use “Social Networks” to work, to keep in touch with friends or simply for fun. This tendency to use these ”communication platforms” has made social networks the undisputed masters of media communication on the web in recent years. Users can share information, take care of their interpersonal relationships or create new ones, can advertise their business or even set up real marketing campaigns. There are different types of social networks like professional or entertainment and specific categories: animals, sports, music, etc. Unfortunately, it is not always good to expose your personal data on social media. A news published on the web, a post on a social network, an inappropriate comment on a chat of a facebook or a ”whatsapp” group are able to easily reach an unspecified number of people and can, however, be quite dangerous whenever the the subject of the diffused message has a disparaging and defamatory nature towards its recipient. In recent years the issue of defamation through the use of social networks has been the subject of extensive debates because it is one of the criminal offenses that are most commonly used. Thanks to anonymity, the web induces the most impudent (called haters or keyboard lions) to offenses and insults of all kinds .

The cases of insult and defamation on these social networks are increasing and it is necessary to be ready to react in the appropriate forums. Defamation or offense that occurs on social media is punishable and their certified acquisition can become evidence in criminal or civil court proceedings.

2.2 Overview on Social Network Platforms

Social networks, born in the late nineties, allow users to create an appropriate user profile, to organize a list of people to keep in touch with, to publish their own stream of updates and also access that one of others. A social network [34] is a service offered through the Internet, typically usable in a completely free way through the web or by specific applications for mobile devices, whose purpose is to facilitate the management of social relationships by allowing communication and sharing of digital content. Most social networks have common characteristics that can be identified in three main elements:

- the creation of a personal profile (public or semi-public);
- the creation of a list of friends;
- the exploration of one's own network and that one of friends.

Once the registration phase is completed, step in which we are asked to provide an e-mail address and a password, we move on to the creation and management of a personal profile of the user through a series of questions (on the city of birth, on the place where we live, on the school we attend or have attended, on the job, on personal interests, hobbies and more). This page contains general information about the user, which can be supported by images or photos, videos and a short self-description. Then we move on to the creation of a list of friends. The list of friends is expanded with the help of computer programs of the social network which, with reference to the answers we gave in the user profile compilation phase, suggest friends we met in real life and ”potential new friends”, selecting from the registered users, people who have characteristics corresponding to our indications. Another feature of social networks is the ability to explore the profiles of friends who are part of our friends list and those who are part of our friends' friends (even if they have not made friends directly with us); you can visit the personal pages of users (friends), observe their favorite activities, musical tastes, etc. and of course interact directly with people we don't know.

2.2.1 How many are the Social Media Sites and Apps?

The world of social media continues to maintain a great power of attraction, in the last year the number of users has grown again by 17%. According to the site [5], each person who frequents social networks is registered on an average of seven sites. The existing social networks are not limited only to Twitter, Facebook, LinkedIn and Blog. But how many social networks are there in the world? Wikipedia [6] lists 206 of them, while [7] has registered 250 ones. In alphabetical order they go from Academia.edu, a site for teachers and researchers that helps to make their work known by sharing scientific publications, up to Greek Zoo.gr frequented by those who want to play online. Some have a few thousand members and are dedicated to individual passions, such as books or cinema, or to communities of people who share particular situations. The social network founded by Mark Zuckerberg remains the most popular and today has more than 2 billion and 100 million users.

1. **Vero** states that its mission is to make available online the natural bounds that already exist among people. Vero has about 3 million users as of March 2018.

2. **Desentric** is a very new social network. It is similar to Twitter and it has some interesting features such as democratic algorithm, transparent functions, mutual growth principle and incentive for founding members.

3. **Befilo** is a new social network where everybody is automatically friend with everybody. Users just join the network and automatically friend with all other members.

4. **Opportunity** is a business social network with a matchmaking algorithm to connect professionals for business and relationship opportunities. It has several million users around the world.

5. **Zoimas** is an anti-addiction social network which keeps users online as little as possible. Users can login only once in 12 hours, be online max 15 minutes, post only once each login and have max 150 friends.

6. **Facebook** is still the largest social network in the world. It is said to have around 2 billion monthly users as of December 2017.

7. **Twitter** has about 320 million users, who can post tweets limited to 280 characters.

8. **Instagram** is a photo and video sharing social network. It's part of Facebook and has around 800 million users as of January 2018.

9. **Pinterest** is a social network where content is added in the form of pins and it has about 200 million users as of January 2018.

10. **LinkedIn** is a social network platform that is mainly used by business professionals. As a trademark of Microsoft, LinkedIn has approximately 500 million users as of January 2018.

11. **Okuna** is a very new social network which started in 2019. It describes itself as privacy-friendly, open and accountable.

12. **Reddit** is a content sharing social network with over 500 million monthly visits. Text posts or direct links can be shared on the site and voted by members to determine popularity.

13. **Gab** is an ad-free social network that allows its users to read and write messages of up to 300 characters, called gabs. It has roughly 200,000 users.

14. **Mastodon** is a new Social network, something between Instagram and Twitter. It is becoming popular fast and adding new features that don't exist on the other two.

15. **Ello** is a global community Social network that brings together artists and creators.

16. **Woddal** is social network that allows people with similar interests to come together and share information, photos and videos.

17. **Hello** is a social network where users can join and create communities with same interests.

18. **MeWe** describes itself as the Generation-Next social network. It offers its users a good deal of autnomy over their presence and security on the network.

19. **Raftr** is an events and communication platform for college students with an exclusive social network for each school.

20. **OneWay** is an alternative social network that focuses very much on political freedom, the right of free speech and liberty of conscience.

21. **EyeEm** is a visual content based social network. It is considered as an alternative to Instagram. But it has a more authentic content provided by 22 million users.

22. **Care2** is a social network that connects activists from around the world to primarily discuss political and environmental issues. The site has about 40 million users.

23. **GirlsAskGuys** is a opposite-sex based social network platform in which opposite sexes ask and answer each other questions.

24. **Houseparty** is a video based social network where users hang out as a group through a FaceTime functionality.

25. **Gentlemint** is a Social network dedicated only to men to talk about and share men-related things.

26. **Canoodle** is a dating Social network that brings together people with same interests.

27. **Bucketlist** is a social network where users can set goals and interact with other users with similar goals.

28. **Brainly** is a large educational social network by students for students. It has 80 million users.

29. **hi5** is one of the oldest Social networks that is popular in Asia, eastern Europe and African countries. It has about 80 million members.

30. **Tiktok** is a short form mobile video share social network. It has about 150 million daily active users and is very popular in China and Asian countries.

31. **NapSack** is a family and friends based Social network. Its motto is to make social networking easy and simple.

32. **DeviantArt** is an art-sharing network with over 38 million registered members.

33. **Flickr** is a photo and video sharing social network that supports tens of millions of members and over 10 billion photos.

34. **My Muslim Friends Book** is a social network for connecting Muslims across 175 countries. The site currently has roughly 500,000 members.

35. **BlackPlanet** is a social network for African Americans that focuses on dating, showcasing talent, and chatting and blogging. The site has around 20 million members.

36. **Steemit** could be seen as a combination of Quora and Reddit. Users can publish their posts on Steemit and based on the upvotes, they can receive Steem crypto tokens. It has about 10 million users.

37. **Amino** is a mobile social network that has large online communities about almost any topic.

38. **InLinx** is considered as one of the new trends in social networking. It brings together many different features sucha as; interaction among families, friends and collogues, dating, making new friends and digital marketing.

39. **Twoo** is a social discovery platform that allows its 181 million members to create profiles, upload pictures, and chat with other users.

40. **MeetVibe** is a mobile phone based social network which allows its users to view the profiles of people nearby at that moment.

41. **Spinchat** is a social network where you can meet new people and play games with them.

42. **Soup** is a social network where users are encouraged to share cool stuff. It has about 4 million members.

43. **Diaspora** is a decentralized Social network that offers many different ways to share your posts and to socialize with other users.

44. **Yooco** is a social network platform which allows its users to create their own social networks.

45. **Meetup** is a social network that facilitates a group of people to meet in person around a specific topic or theme. It has roughly 32 million users.

46. **Badoo** is one of the world's most widely used dating networks. It has over 360 million registered users.

47. **Tagged** is a social network for making new friends. The site has about 20 million unique visitors globally.

48. **VK** is like Facebook but more popular in Russia and neighboring countries with over 400 million users.

49. **Tumblr** is a blogging network with over 350 million blogs and over 500 million users. The social network supports both web and mobile.

50. **Hub Culture** is a Social network that allows its members to create connections in the physical and digital world.

51. **Cookpad** is a recipe social network and a community platform for people to share recipe ideas and cooking tips. It has about 100 million monthly users from 70 countries.

52. **Streetbank** is a sharing and neighbouring social network. Users basically do three things: Give things away, share things and share skills. It has about 20 thousand users from around the world.

53. **Happn** is a social network that brings together people who are at the same place, same time. It has about 70 million users.

54. **Snapchat** is a mainly audio-visual content network with around 200 million users as of January 2018.

55. **MeetMe** focuses on helping users discover new people to chat with on mobile devices. It has over 2.5 million daily active users.

56. **WeGather** is a new social network where users can meet to discuss the topics thy like.

57. **Peanut** is a social network which helps mothers connect with other mothers. Users Exchange emotional support, stories, advice and many other similar interests.

58. **Peloton** is a fitness and cycling based Social network. Members get live advice from instructors while cycling outdoor or at home. It is popular in western countries.

59. **WeChat** is a mobile-messaging social network with almost 1 billion monthly active users who are primarily from China. But WeChat also offers an English, international version. It has rich functionality from chatting to shopping with users even buying homes on the app.

60. **FicWad** is a social network where users can create stories and their stories shown on search results.

61. **Digg** is a news based social network. It curates news, articles and videos from other major media sources. The users can digg a story or a piece of news to be seen by other users.

62. **Anchor** is a social network which allows its members to easily send podcasts and distribute them widely with just one click.

63. **Discord** is a social network that offers free and secure voice and text chat opportunity fort he gamers online. It is considered as an alternative to Skype and TeamSpeak.

64. **Blab** is a social network which offers the freedom of sharing anonymously. Users can share their secrets, confessions, fears, and funny stories anonymously.

65. **Bebee** describes itself as a personal branding social network. Users unite their personal and professional lives in one profile and market themselves to employers, clients and vendors.

66. **Signal** is a social media app which allows users to send high-quality group, text, voice, video, document, and picture messages anywhere in the world without SMS or MMS fees.

67. **Douban** is a very large Chinese social network which brings together book and film lovers and music fans.

68. **Skyrock** is primarily a French social network that offers blogging capabilities to its members. It has a few million members.

69. **Irc-Galleria** is a social network in Finland. It has about 500 thousands members and is open mainly to Finnish speaking people.

70. **Wer-Kennt-Wen** is a community Social network in German language.

71. **Mamba** is a dating social network which is most popular in Russia and many other countries in the world.

72. **Copains d'Avant** is the number one social network used in France.

73. **Ameba** is one of the largest social networks in Japanese.

74. **Miarroba** is a Spain-based social network which allows its users to share different types of content.

75. **Fc2** is the third biggest social network in Japan. It is also available in many other languages.

76. **Weibo** is a large Social network in China which has about 300 million members.

77. **Spaces** is a Social network mainly popular in Russian-speaking countries.

78. **StudiVZ** is a Social network dedicated to students and teachers in German-speaking countries.

79. **Taringa!** is a Social network that is very popular in Argentina and other Spanish-speaking countries.

80. **Trombi** is a french social network where members find and connect old friends. It has more than 9 million users.

81. **Zoo.gr** is a social network for Greek people to meet and connect.

82. **Hatena** is a Japanese social network known with its bookmarking feature. Users interact via the urls they have shared.

83. **LiveInternet** is one of the largest social networks in Russia. Its members are estimated around 25 million.

84. **Rediff** is an India-based social network and portal similar to pinterest.

85. **Qzone** is one of the largest Social Networks in China. It has 480 million members and is only in Chinese. Also 9th largest website in the world.

86. **Plurk** is a Social network especially popular in taiwan which alows its users to create and share content in short plurks.

87. **Mixi** is a popular Social network in Japan. It has about 25 million members.

88. **Live Journal** is a Social network that is very popular in Russian-speaking countries.

89. **Nasza Klasa** is a very popular Social network in Poland.

90. **Zing** is the largest social network in Vietnam. It has about 7 million members and is considered larger than local facebook.

91. **Cyworld** is a South Korean Social Networking website. İt has around 20 million members and is only in Korean langauge.

92. **Cloob** is a social network that mainly serves Iran and Farsi speaking countries.

93. **Nextdoor** is a social network that connects neighbors by sharing upcoming events and other neighborhood activities. Over 150,000 neighborhoods in the U.S. use Nextdoor.

94. **Tuenti** is a Social network dedicated to university and high school students. It has about 12 million members and is especially popular in Spanish-speaking countries.

95. **Gaia** is an anime-theme social network and forums-based website. It has over 25 million registered users.

96. **Wize** is the largest German-language social network for Best Agers (those in their 40s and up) where they meet and make new friends.

97. **Odnoklassniki** is a very popular Social network in Russian-speaking countries and former Soviet Union countries.

98. **Partyflock** is a Dutch Social network that brings together members interested in house music and general electronic music.

99. **Renren** is another large Chinese Social network with about 200 million members, especially popular among university students.

100. **Wanelo** is an online shopping social network. Users post and sell any kind of products under the guidance of the network.

101. **Internations** is a social network that connects expats across 390 cities worldwide. It has almost 3 million users.

102. **Cross** is a social network that shares Christian content to its 650,000 members.

103. **AngelList** is a social network platform mainly used by new investors and startup entrepreneur.

104. **Xing** is a career-oriented social network that is used by consumers and businesses. Xing supports closed groups to enable a private and secure network within an enterprise.

105. **Storia** is a Social network where users can create and share their stories. The network has around 10 million members.

106. **ASmallWorld** is a paid social network that can only be joined based on an invitation by a member. The site focuses on luxury travel and building social connections, Its membership is capped at 250,000.

107. **ReverbNation** is a social network for musicians to help them manage their careers and find new opportunities. The site has about 4 million musicians as members.

108. **SoundCloud** is an online audio distribution platform that enables its users to upload, record, promote, and share their originally created sounds. The service has more than 150 million unique listeners every month.

109. **Medium** is probably world's largest social network for reading and writing. It has around 60 million users.

110. **eToro** is a worldwide Social investment network bringing Social traders together.

111. **HR** is a Social network for Human Resources professionals worldwide.

112. **Influenster** is a Social network for revies and sampling of new products online. It has about 1 million members.

113. **Evernote** is an international social network which connects business professionals. It has about 15 million users.

114. **23snaps** is a social network for families to share the online photos of whole family in private albums.

115. **Flixster** is a site for discovering new movies, learning about movies, and meeting others with similar tastes in movies.

116. **Minds** is a social network that allows its users to create channels on a variety of topics and also rewards users for their online activity. It promotes freedom and privacy on the Internet and has over 2 million members.

117. **Photobucket** is a photo and video hosting site that has over ten billion photos and over 100 million members.

118. **CaringBridge** is a social network for people facing various medical conditions, hospitalization, medical treatment, and recovery from a significant accident, illness, injury, or procedure.

119. **About.me** is mainly serving freelancers and entrepreneurs who want to increase number of their clients. It has about 5 million members.

120. **CafeMom** is a site for mothers and mothers-to-be. It has over 8 million monthly unique visits.

121. **Ravelry** is a social network for knitting, crocheting, spinning, and weaving. The site has over 7 million registered users.

122. **Slashdot** is a social network where users can add their news and articles to be commented by other users.

123. **Playlist** is a musical social network. It allows it users to create unlimited music lists and share with friends. It has about 500 thousand users.

124. **Airtime** is a new video sharing social network. It allows member to gather for group chat and video watch.

125. **Cellufun** is a gaming community with over 2 million members that can be accessed using any mobile device.

126. **MocoSpace** is social gaming site with over 2 million users and over 1 billion monthly page views.

127. **Zynga** offers multiple games that are played by millions of daily users. Popular titles are Farmville, Draw Something, and Zynga Poker.

128. **Habbo** is a social gaming company for teenagers. It has more than 5 million unique monthly visitors. The network operates nine sites for users in different countries.

129. **Crunchyroll** is a Social network for those who like anime, cartoons and the like.

130. **Elftown** is a Social network that has a community interested in fantasy and sci-fi arts and literature. It has about 200,000 members.

131. **Ryze** is a social and business network. Members can grow their business, build career, find a job, make sales or just keep in touch with old and new friends. The network has about 1 million members.

132. **Rooster Teeth** is a Social network dedicated to online games, webseries, music and anime.

133. **Twitch** is a social network dedicated to online games.

134. **Hyves** is the most popular social network in Holland with about 10 million users.

135. **Fishbrain** is a social network especially for those who love fishing. It has about 2 million members.

136. **TalkBizNow** is a social network for business people. Members can meet and do business together on the network.

137. **OpenDiary** is one of the oldest social networks (founded in 1998). It allows members to have online diaries. It has about 5 million members.

138. **Athlinks** is a social network which brings together pople interested in outdoor activities like running and swimming.

139. **RallyPoint** is a social network with over 1 million current and former members of the US military gather to discuss military life, share information, and exchange stories.

140. **BeMyEyes** is designed to help blind people to solve everyday problems. The app connects the vision-impaired with fully-sighted users via video chat to show the situation and get help. In a sense, the sighted person lends their eyes to blind people.

141. **Kik** is a new mobile social media app for online chat and instant messaging.

142. **GetJealus** is a social network where members share travel related content.

143. **TravellersPoint** is an online travel community network where users share their travel exper,ences.

144. **Hospitality Club** is a Social network that brings together hosts and guests, travelers and locals to find free accomodation worldwide.

145. **Gapyear** is Social network that brings together travellers worldwide.

146. **Tournac** is a social network for travelers that connects people traveling to the same location.

147. **CouchSurfing** provides a platform for members to stay as a guest at someone's home, host travelers, meet other members, or join an event. The site has roughly 15 million members.

148. **YY** is a another large social network of China with 122 million users. It allows its users to post videos to a group which can be watched by hundred thousands of members.

149. **Glocals** is a social network created in Switzerland for the expatriate community. It allows the members to meet, organize activities, and share information.

150. **Nexopia** is a Canadian social network that allows its members to create forums on any topic and have discussions within those forums. The site has over 1 million users.

151. **Classmates** connects people with their high school friends in the U.S. and also allows for the uploading of high school yearbooks. Members can also plan their high school reunions.

152. **Tingles** is a video social network based on ASMR (Autonomous Sensory Meridian Response) which allows soft-sounded videos to trigger positive, euphoric feelings. It has several million users.

153. **Quora** is a question-answer based social network platform where users ask-answer questions. It has about 200 million users as of January 2018.

154. **Ask** is another large question and answer social network similar to Quora. It has about 160 million users.

155. **ProductHunt** is a social networking website which gives priority to content about new products.

156. **Untappd** is a mobile social network that allows it members to rate the beer they are consuming, earn badges, share pictures of their beers, review tap lists from nearby venues, and see what beers their friends are drinking. The site has roughly 3 million members.

157. **Doximity** is a social network for U.S. clinicians. It has over 800,000 members.

158. **WriteAPrisoner** is a US-Florida based social network bringing together users and children impacted by crimes.

159. **Xt3** is a Catholic social network founded for the youth in Australia. It has about 70,000 members.

160. **Altervista** is an Italian social network where users can create websites free.It has about 2,5 million users.

161. **MixCloud** is a social network where users can listen to DJs, create and share their lists with other users.

162. **BranchOut** is a Social network for those looking for jobs and future carieers.

163. **Koofers** is another educational social network with about 2 million members. It is popular among college students.

164. **Edmodo** is an educational Social network where teachers, students and parents connect. It has about 50 million users.

165. **PatientsLikeMe** is a Social network for patients connecting patients with similar illnesses to exchange information.

166. **DailyStrength** is a medical and support-community based Social network with about 43 million members.

167. **MyHeritage** is an online genealogy network that enables users to create family trees, upload and browse photos, and search billions of global historical records. The site has 80 million users worldwide.

168. **23andMe** is a DNA analysis company that connects its customers with their relatives based on a DNA analysis. It also identifies if the person is likely to have any health-related issues based on DNA analysis.

169. **Ancestry** is in the business of finding your ancestors — i.e. building genealogy networks. The site has roughly 2 million paying members.

170. **Geni** is a Social network that allows its users to create their family tree and invite other relatives to join. It has about 180 million users.

171. **Bandcamp** is a social network that connects musicians and artists.

172. **VampireFreaks** is a community for gothic-industrial subcultures that has millions of members. The site is also used for dating.

173. **Tinder** is a location-based dating mobile app that is used by over 50 million users.

174. **Crokes** is a community or social network for authors. It is similar to Twitter, but limits posts to 300 characters.

175. **Goodreads** is a social network for book lovers, who can recommend books and see what their friends are reading, among other features. The site is owned by Amazon and has tens of millions of members.

176. **Academia** is a social networking website for academics. The platform can be used to share papers, monitor their impact, and follow the research in a particular field. The site has over 55 million users.

177. **Library Thing** is a Social network dedicated to books and book reader community.

178. **Listography** is a Social network with lists and autobiography.

179. **Bibsonomy** is a Social network where members can organize scientific work, researches, collect publications, and contact likeminded colleagues and researchers.

180. **ResearchGate** is a social network where researchers and scientists can meet, discuss and exchange their knowledge.

181. **Wattpad** is one of the largest literature based social netwoks where readers and authors connect. It has about 65 million users.

182. **Anobii** is a social network where readers can connect and exchange ideas about books.

183. **Scribd** is a large social reading network where members can read books, audio books and magazines.

184. **Grindr** is a mobile social network for gay, bi, trans, and queer people to meet and connect.

185. **OutEverywhere** is another gay social network where LGBT people meet and connect.

186. **Gays** is a Social network for the LGBT community. It has more than 100,000 members.

187. **Letterboxd** is a social network where users like, review and share content about films.

188. **Kroogi** is a social network in Russian and English which brings together artists,musicians and painters. It has about 100,000 users.

189. **Stage32** is a Social network and educational website for people in TV, cinema and film industry.

190. **Film Affinity** is a Social network bringing together people with smilar liking of movies and TV series.

191. **Filmow** is a Brazil based Social network which allows its users to list, rate and recommend the films they watch.

192. **Telfie** is a Social network for entertainment.

193. **Indaba Music** is a Social network for music community worldwide.

194. **Mubi** is a subscription based Social network for cinema community.

195. **Lingualoe** is a language learning-teaching social network. It has about 4 million members.

196. **Amikumu** is another language learning social network. It helps the user to find the nearby speakers or learners of the same language.

197. **Hellolingo** is a Social network dedicated to teaching and learning foreign languages.

198. **Italki** makes connections between language learners and language teachers to help learning new languages. The site has more than 1 million students.

199. **English, baby!** is a social network and online curriculum for learning conversational English and slang. The service is used by more than 1.6 million members.

200. **Busuu** is a language-learning social network. The site connects learners to speakers of the native language to make the learning process easier.

201. **WhatsApp** is an instant-messaging social network platform that is mainly used on smartphones. It has recently been bought by Facebook and is estimated to have about 1 billion users as of January 2018.

202. **Messenger** is another instant-messaging social network platform which functions inside Facebook. Its users are estimated around 1.2 billion as of January 2018.

203. **Skype** is an instant messaging platform that enables communication using text, voice, and video. It has over 300 million active monthly users and is now part of Microsoft.

204. **Viber** is also a communication social network like Skype that allows text, voice, and video messaging. It has over 800 million users

205. **Line** is an instant messaging social network that is popular in Japan but also supports English and other languages. It has over 600 million users worldwide.

206. **Telegram** is a cloud-based instant messaging service that has over 100 million active monthly users.

207. **Stumbleupon** focuses on content discovery for its users. It is offered as a browser toolbar in all popular browsers.

208. **YouTube** is the world's leading video sharing network that enables its users to upload, view, and share videos. It serves billions of videos daily.

209. **Vimeo** is a video based Social network very much like YouTube, but with more and different features and functions. It has 70 million members and 240 million monthyly viewers.

210. **FunnyOrDie** is a comedy video network that allows users to upload, share, and rate videos. The videos often feature celebrities. The network has hundreds of millions of viewers.

211. **Tout** is a video network that helps businesses grow online video revenue and drive deeper engagement with viewers. It has 85 million unique monthly viewers.

212. **Last Fm** is a music discovery and recommendation network that also shares what friends on the network are listening to. The site has tens of millions of users and over 12 million music tracks.

213. **Howcast** is a social network similar to Youtube where users can upload high quality how-to video content.

214. **Bigo** is a live streaming social network where users can showcase their talents and meet other members. It is very popular in Singapore, Thailand, Japan and India and has about 40 million members.

215. **Bitchute** is a video hosting social network, like YouTube. Its policies, and especially monetization policy, is considered to be less strict tha YouToube.

216. **Snapfish** is a photo sharing social network where the members can benefit from unlimited storage space for their photos. The site has tens of millions of members.

217. **Blogster** is a social network and blog which allows its users to create blogs and interact with each other. It has about 1,5 million users.

218. **Shutterfly** is a photo sharing site that allows its 2 million members to use the photos to create personalized gifts, such as mugs and t-shirts.

219. **500px** is a Canadian photo sharing social network with over 1.5 million active members.

220. **Dronestagram** is focused around sharing photos that have been taken using drones. It claims to be “the Instagram for drone photography,” with more than 30,000 members.

221. **Fotki** is available in 240 countries. It has more than 1.6 million members and 1 billion photos. The site was started in Estonia.

222. **Fotolog** is a photo-blogging site with over 20 million unique visitors.

223. **Imgur** is a photo-sharing site where members can vote (and rank) photos. The site has hundreds of millions of images.

224. **Pixabay** shares high quality photos from its members. The site has over 1.1 million images and videos.

225. **WeHeartIt** is a social network for sharing inspiring images. The site has over 45 million members.

226. **PlentyofFish** is a dating social network that is free to use but also offers some premium services. It has over 100 million registered members.

227. **Viadeo** is a social network for business owners, entrepreneurs, and managers — mostly in Europe. It has about 50 million members.

228. **Foursquare** provides personalized recommendations based on a user's location and previous purchases. The service has tens of millions of users and is growing rapidly in the enterprise space.

229. **Yummly** is a social network dedicated to food recipes and cooking.

230. **Pinboard** is a paid social network that allows sharing of bookmarks. The users can benefit from an ad-free experience on this site.

231. **Daniweb** is a social network where IT people meet and discuss. It has about 2 million users.

232. **Swarm** is a new social network by Foursquare which helps people to remember all the places and venues they have been to.

233. **Yelp** is a restaurant review and home services site that has social features to share photos, write reviews, and see activities of friends.

234. **Threema** is a text, video and voice messaging social network. Members are able to use it without providing their email or mobile phone info.

235. **Ning** is a social network which allows its users to create a social website and get it monetized.

236. **Stack Exchange** is a question-answer based social network similar to Quora.

237. **Jsfiddle** is a social network where users test and showcase their HTML, CSS and JavaScript codes.

238. **Houzz** is a social network where users connect and share design and decoration related content.

239. **Dribble** is a social network which mainly allows designers to connect and share ideas.

240. **Disqus** is a social network which allows its members to build online audience around their content or website.

241. **Bitbucket** is a social network whare users can share script codes and ideas about coding.

242. **Slack** is a social network that brings together team members on certain works and projects.

243. **SlideServe** is a large social network where users can upload and share their slides and Powerpoint presentetaions.

244. **The Dots** was created in 2014 in Egland. Its mission is to help the no-collar professionals to connect and cooperate for creative projects.

245. **Zotero** is a social network and a free software which functions as an assistant for web research.

246. **Webnode** is a social network that brings together users on basis of free website building. It has 30 million users.

247. **Caffeine** is a game social network which allows its users to broadcast the video of the game they are playing.

248. **Brave is** a social network for website builders, email marketers and the like. It has 15 million users.

249. **GoFundMe** is a fundraising network that can be used to raise money for most any cause.

250. **Kickstarter** is a social funding platform where people can pitch their products or product ideas to get funding. The site has had almost 10 million backers.

2.3 Using Social Media in a Criminal Procedure

In Italian law, defamation is the crime provided for by art. 595 of the Criminal Code, which reads: ”Anyone who, apart from the cases indicated in the previous article, by communicating with more people, offends the reputation of others” and ”is punished with imprisonment of up to one year or a fine of up to euro 1,032.00 ”. Law 48/2008 has introduced into our legal system a series of new offenses generically classified as computer crimes, but has not added anything for the possibility of configuring the defamation crime through computer networks or telematics. In the typical crime of defamation, the principal legal asset is the reputation and its structural elements are: the offense to the reputation of others that is an injury to personal, moral, social, professional qualities, etc. of an individual; communication with several people, where the expression ”several people” must certainly be understood as ”at least two people”; the absence of the offended person. It is not difficult to argue, however, that the offense referred to in art. 595 of the Criminal Code, is sufficiently generic to also include all those offensive behaviors that are carried out through computer networks and modern communication techniques.

2.4 Investigate on Social Media Crime

Forensic analysis of social networks is the method used by investigators to identify and prosecute dangerous subjects present in a social network service.

The acquisition and collection of evidence of defamation is a key element, but it is necessary to certify and verify the integrity and authenticity of the collected evidence. The authentication of the page or profile that has carried out the defamation, insult or slander can be followed by a Notary equipped or by a forensic computer expert who acquires the pages with the messages defamatory or abusive. The digital appraisal aimed at documenting the defamation and the offense or injury occurred on the Internet through computerized evidence that can be extended through OSINT investigations and searches also to the acquisition of data relating to the owners or users of the profiles, groups or pages on where defamatory messages are published.

2.4.1 OSINT e SOCMINT

The Open Source INTelligence, acronym OSINT is the activity of collecting information by consulting publicly accessible sources. OSINT sources are distinguished from other forms of intelligence because they must be legally accessible to the public without violating any copyright or privacy laws. Indeed, OSINT includes all sources of information accessible to the public. This information is available online or offline, let's see some examples:

- Access to the Internet, which includes forums, blogs, social networking sites, video sharing sites, wikis, Whois records of registered domain names, metadata and digital files, geolocation data, IP addresses, people's search engines and everything that can be found online.

- Traditional mass media (TV, radio, journals, book).
- Specialized journals, academic publications, dissertations, conference proceedings, company profiles, annual reports, company news, employee profiles and resumes.

There are organizations specializing in OSINT services. Some of them are based on government services others are private companies that offer their services to various entities such as government agencies and commercial companies on a subscription basis; among the best known: government bodies, international organizations, military agencies, but also companies whose power is information.Uno dei passaggi principali durante l'attività di OSINT è identificare gli indirizzi IP e i sottodomini associati al bersaglio.

The term of Social Media Intelligence (SOCMINT) indicates a set of techniques and technologies that allow private or public agencies to monitor social media platforms. SOCMINT's activities concern the monitoring of content, such as messages or photos posted, and any other kind of data produced during an activity session on social media. Such information, whether private or public, involves interactions between people, between people and groups or between different groups. The methods used to analyze the data produced through social networks are different: they may also include the manual correction of content, public or private, or of entire pages; o reviewing the results of some research or some questions; o modification of activities or content posted by the user; or scraping, which translated means scraping and which consists of extracting the content of a web page and duplicating it in a way accessible to those involved in social media intelligence. Clearly, SOCMINT's activity includes a series of procedures to collect, store, and analyze the data produced on social media, data that are subsequently translated into analyzes and trends. The term Social Media Intelligence is sometimes replaced by the equivalent Open Source Intelligence (OSINT), although there is a substantial difference between the two activities': while the OSINT analyzes only public data, such as articles, sites and blogs, SOCMINT analyzes both those public and private ones, i.e. messages and chats.

2.4.2 Identify profile, page or group with defamatory content

In order to perform a correct analysis on a profile, page or group, it is necessary to identify the ID code that uniquely identifies it. The profile name can in fact be changed by the owner, as well as the address that appears in the browser's URL bar. To locate the ID code of the profile or page from which the defamation comes, you can use a site such as Find My FB ID, by pasting the profile or page address in the text field and pressing the ”Find numeric ID” button. Once you have entered the address of the profile or page where the defamation is present, you will get a number to copy or print, to ”freeze” the unique identifier that will allow you to find the profile or page even in the event of a name change or URL and to ask the Judicial Authority for any log files or defamatory content. If it is not possible to use online sites that identify the ID, it is advisable to save the page or profile on which the defamation was detected. Within the page code, you will find two items containing the ID codes searched: ”pageID”;(for Facebook pages) and ”profile_id” (for profiles).

2.4.3 Find the unique reference of the defamatory post or comment

Once the User ID of the owner of the profile from which the defamation occurred or the Page ID of the page containing the defamatory text has been established, it is also necessary to ”freeze” the post or comment itself, including the date, to then use it as IT proof of the defamation and allow the IT forensic consultants who will be hired to carry out an IT expertise. The address or URL that identifies the post itself will be of the following type: www.facebook.com/profile.name/posts/10213357451991856 . To identify a specific comment, by clicking on the date and time under the comment itself, after the "Like" link, the post will be opened in a new page with the comment highlighted, a URL of the following type: www.facebook.com/profile.name/posts/10213357451991856?comment_id=10213357955884453

The first code, highlighted in bold, is the ID code of the post, while the second one is the "comment_id", that is the unique identifier of the defamatory comment.

2.4.4 Comments removed

In the event that the defamation continues with further messages and insults in a thread or in a post, especially if they are then promptly removed, in the case of some social networks it may be possible to activate the setting that allows you to "follow a post" so as to receive a mail at each update. You will thus receive emails for each new comment to the post, which can be used by the IT expert to certify or trace the comments even if they should be removed shortly after publication.

2.4.5 "Freeze" a digital proof of defamation

It is always important to make a certified copy of a profile or page containing defamatory messages.

However, it is possible (also to protect oneself in the event of cancellation) to begin the "crystallization" phase using some precautions, such as the free FAW (Forensic Acquisition of Websites) software that allows for the forensic acquisition of web pages or social profiles network with some guarantees on the originality of the acquired data. There are also web services that allow you to download an authentic copy of pages or posts as long as they are public and not private, such as Perma.cc or Archive.is that allow you to create a copy of an Internet page on a third server, carried out by a third party, a strategic activity in particular in the event that the defamatory messages are modified or removed.

2.5 Forensic Analysis on Social Media Applications

In the recent years, Social Media Applications received attention from many forensic researchers, because of their exponential growth, due to their ease of use and efficiency reaching out to people, allow the development of many malicious activities and serious cybercrime [100]. In 2012 Al Mutawa et al. [21] focus their attention on mobile device analyzing forensic artifacts of several Social Media apps on various mobile platforms: MySpace, Twitter and Facebook each on Blackberry phone, iPhone (iOS) and Android. In 2013, M. Baca et al [26] conduct an analysis of Facebook artifacts in internet and were able to find significant evidence traces related to Facebook activity. Other research based on the analysis of WhatsApp, Viber and Skype artifacts was carried out [93] [127] [22]. In 2015 Walnycky et al. [135] conduct a network and device forensic analysis of twenty android social messaging apps to explore digital evidence strictly limited to messaging service only. A forensic analysis of three social media apps (Facebook, Viber and Skype) in windows 10 was carried out by Majeed et al. [94]. They explored and examined the potential locations of storage finding interesting artifacts for all three applications in plain text. In 2017 Yusoff et al. [145] conduct an investigation and analysis of social media and instant messaging services focus on residual remnants of forensics value in FireFox OS. They examined three social media services (Facebook, Twitter and Google+) as well as three instant messaging services (Telegram, OpenWapp and Line). A very interesting research focused on authorship attribution for Social Media Forensics was conduct by Rocha et al. [118]. Their research is based on the fact that all authors possess peculiarities of habit that influence the form and content of their written works. These characteristics can often be quantified and measured using machine learning algorithms. Rocha et al. provide a comprehensive review of the methods of authorship attribution that can be applied to the problem of social media forensics. Further, they examine emerging supervised learning based methods that are effective for small sample sizes, and provide step-by-step explanations for several scalable approaches as instructional case studies for newcomers to the field. A work on Forensics of Social Network Relationship based on Big Data was carried out in 2020 by Junjing et al. [76]. They expound the forensics mode of social network relationship and the forensics process of mobile phones, and puts forward the forensics method of social network relationship based on Wechat platform, analyses the instance data set, obtains the social network diagram, and intuitively and clearly shows the relationship and intimacy between multiple members. The possibility to extract information about images uploaded on social platform has been exploited in [62] [101] .

2.6 Results

Our analysis has been conducted on the 250 social networks listed by [7]. We have examined the different social networks in order to find the way to identify useful information that can be used in forensic investigations, like User Id, Post Id, Comment Id and any other to retrieve univocally to the defamation author. For each Social Network website, different type of approaches has been adopted to trace these identifiers: an inspection and analysis of the URL of the page or profile of interest, and a "Inspect Element" to examine the source code of the profile or post page; the inspection element analysis, mainly in HTML code, aims to find a script containing the string with the useful information for the investigation (for example an ID Users or others relevant ID) . Our analysis has shown that is not always possible to use both these approaches for every social network examined, or they are not always able to discover interesting information to be used in front of a court in the event of any cases of defamation. In some cases the forensic study has not been possible to conduct; for example when it is not able to extract useful evidences or when the social site has characteristics that do not allow the defamation crime.

We have examined the different social networks in order to find the way to identify useful info like user id, post id, comment id and any other to go back to the defamation author. Typically it is possible to trace these identifiers from the URL of the page or profile of interest, or it can proceed from "inspect element" to examine the HTML code of the profile page; within the HTML code you can find a script containing the essential information. It is possible to identify almost for every social network unique identifiers for the various components present within the Social Networks, useful data for evidential purposes at court in the event of any cases of defamation. In the next we present our analysis result. In some cases the forensic study is not possible, because of different reasons like as limitations from

Italy, closed web site, unreacheable app or because the social site has characteristics that do not allow the defamation crime.

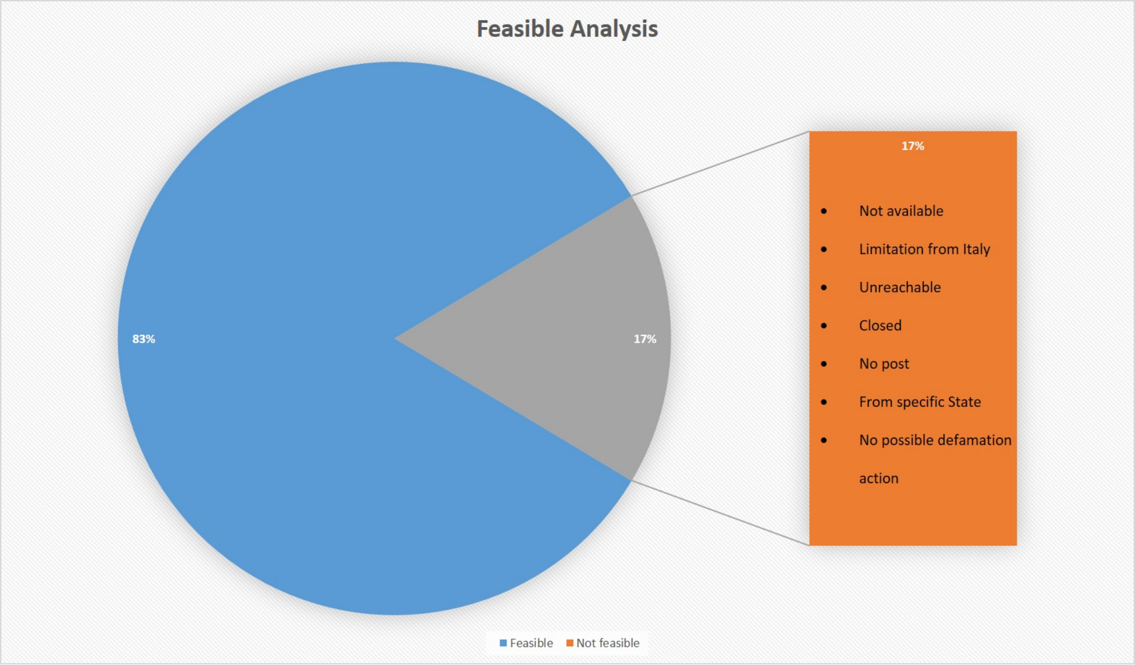


Figure 2.1: Percentage of social network where is possible conducts a forensic analysis. For others is reported the reason because the analysis is not feasible.

Figure 2.1 shows our analysis results conduct on 250 Social Network Platform: in the 83% of the cases it was possible to pull out useful information applying the forensics standard approaches said before. In the remaining 17% of the cases it was not possible conduct a forensic analysis because of different reasons (closed web site, unreachable app or website or because there are not post on the website to analyze). In the Social Media Forensics the main information to detect defamation author are basically obtained from URL analysis or Code Inspection. Focus our attention on the inspectable Social Network Websites, in Figure 2.2 we show how many Social Media is possible to investigate with a URL analysis approach, how many with a Code Inspection approach and how many with both of them.

In the 8% of the analyzed cases it is not possible to proceed with these techniques but

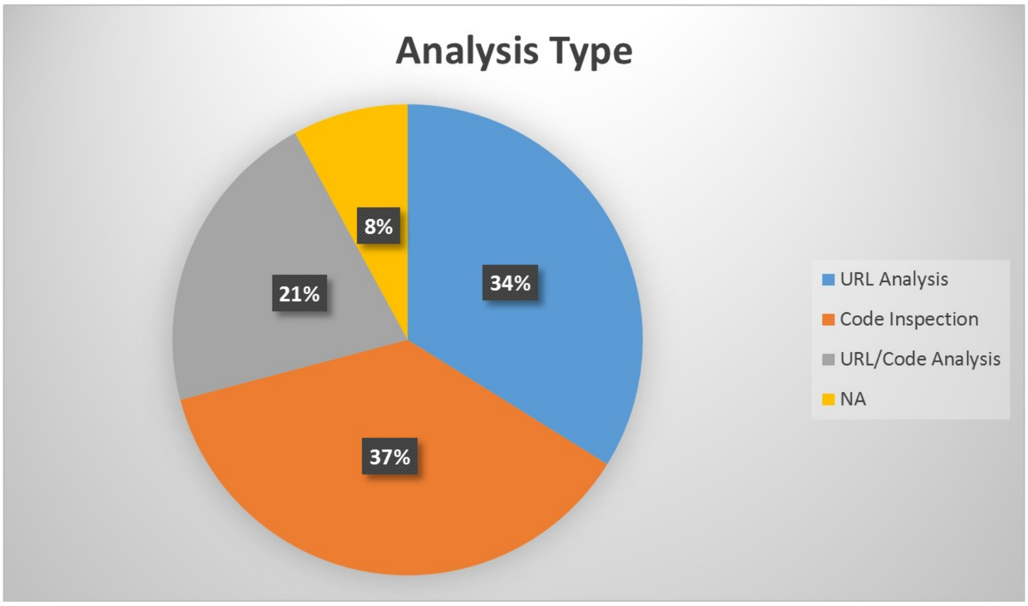


Figure 2.2: Forensic methods applicable to social networks

it is necessary to contact the social network provider to recover valued information. Figure 2.3 shows the number of Social Network where it was possible pull out some evidences such as IDUser, ID Post and ID Comment.

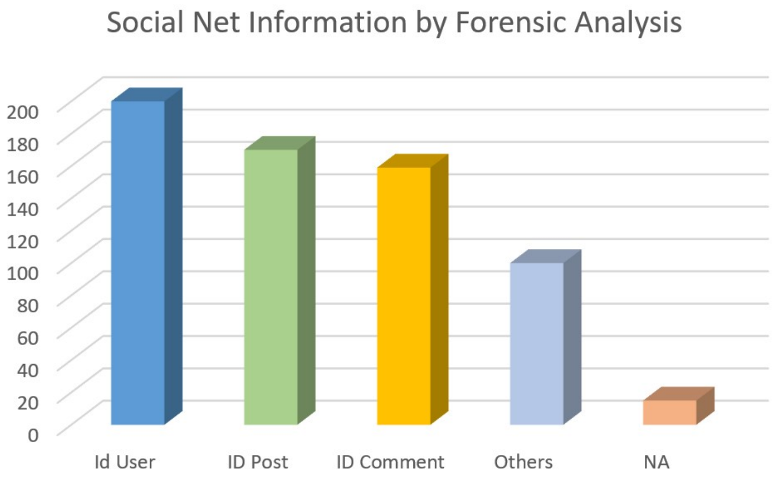


Figure 2.3: Type of social information that is possible pull out from forensic analysis

Moreover, it is possible to find also other numerous information that characterize social network as shown in figure 2.4.

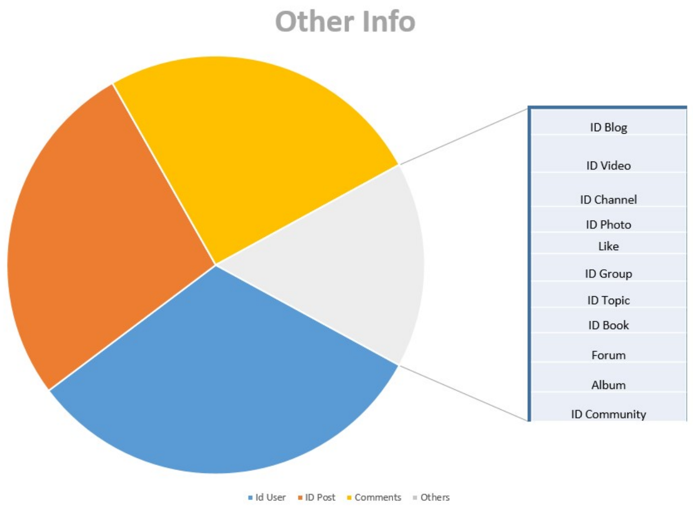


Figure 2.4: Some different social info obtainable from forensic analysis.

Chapter 3

JPEG Compression Algorithm Background

3.1 Lossy and Lossless Compression

Image compression is an extremely important part of modern computing. It addresses the problem of reducing the amount of data required to represent a digital image. It is a process intended to yield a compact representation of an image, thereby reducing the image storage/transmission requirements. Compression is achieved by the removal of one or more of the three basic data redundancies, Coding Redundancy, Interpixel Redundancy, Psychovisual Redundancy . Coding redundancy is present when less than optimal code words are used. Interpixel redundancy results from correlations between the pixels of an image. Psychovisual redundancy is due to data that is ignored by the human visual system. Image compression techniques reduce the number of bits required to represent an image by taking advantage of these redundancies. The objective of compression is to reduce the number of bits as much as possible, while keeping the resolution and the visual quality of the reconstructed image as close to the original image as possible, thus allowing to save disk space and to become easier the transfer of images from one computer to another (which is why image compression has played such as important role in the development of the internet). There has been several image compression techniques developed during the years that we can classify within two main categories: “Lossy” and “Lossless” compression. The “Lossy image compression” can be loose information during the compression to make the image smaller. The “Lossless image compression” on the other hand compresses the image without loosing any information. JPEG image compression algorithm is a “Lossy” format. It provides a very effective way to compress images with minimal loss in quality. This means that the decompressed image is not really the same image as the one you started with. During the years, JPEG compression has evolved tremendously to decrease image file size and at the same time keeping acceptable perceivable quality. JPEG compression allows the user to make a trade-off between image file size and image quality. However, it can be hard to find a good balance between file size and image quality. Imagine that we host a website containing millions of images. If we do not compress the images the website will most likely be very slow. However, if we compress the images too much then the images will look bad. JPEG compression takes advantage of the limitations of the human eyes, namely, the fact that small color changes are perceived less accurately than small changes in brightness.

However, if we analyse images automatically with a machine instead of human eyes, then the small errors that JPEG produces will be a problem, even if the human eyes cannot see them. With "Lossless" compression, the algorithm tries to make the information as compact as possible without loosing any information. So why would we want to use "lossy" compression? The reason is that in the JPEG compression process the information that we throw away is usually something that the human eyes cannot see. Therefore, the image will still have a high quality but the size will be reduced significantly. Something that makes JPEG's lossy compression even more useful is the possibility to adjust the compression parameters. By doing so, we can make a trade-off between image quality and size. If we can accept low quality then we can make the images very small. If it is more important to have an accurate image that is going to be processed by a machine and the size is not a problem, then we simply increase the quality by reducing the compression rate. When compressing a JPEG image, the user is usually given the ability to decide a quality setting that controls the degree of how much the image is going to be compressed. However, there seems to be some confusion regarding these quality settings. The quality range is between 1 and 100. Choosing a quality setting of 50 does not mean that you will keep 50% of the information. The quality scale is actually quite arbitrary. Although the actual implementation of the JPEG algorithm is more difficult than other image format (such as png) and the actual compression of image is expensive computationally and loose some information, the high compression ratios that got attained using the JPEG algorithm easily compensate for the loose if information and the amount of time spent to implement the algorithm.

In the next section we will explore how JPEG algorithm work.

3.2 JPEG Theory

JPEG is an image compression standard used for storing images based on lossy compression format, particularly for those images produced by digital photography. JPEG was created in 1992 [8] and typically achieves 10:1 compression with little perceptible loss in image quality. The term "JPEG" is the acronym for the Joint Photographic Experts Group [9], the joint committee between ISO/IEC JTC 1 and ITU-T Study Group 16 (formerly CCITT) that created and maintains the JPEG. Their goal was to create a standard for image compression. At the time, most computers did not have as much storage as we have today and therefore could not handle largefiles in an efficient way. To solve this, a universal standard was needed for storing image files in a compressed format. Almost 30 years later, the JPEG image file format is still one of the most common image types on the web. JPEG format is quite popular and is used in a number of devices such as digital cameras and is also the format of choice when exchanging large sized images in a bandwidth constrained environment such as the Internet. The reason is that the JPEG format uses a compression technique that allows large image files to be compressed to a much smaller size without loosing visible quality of the image. This makes JPEG images an excellent file format for storing images. The JPEG compression has a simple goal. We want it to take a large file and compress it to a smaller file without loosing too much visible quality to make it easier to transport and/or store. When displaying webpages, text files are much smaller than images and we therefore need to compress the images to achieve acceptable loading times.The problem is finding the right compression rate because there is no good way of knowing how much we should compress an image to retain good perceived quality. Using the same compression settings on two different images can yield different results. There might be no visual loss of quality in the first image while the second image can look very poor for the human eye. The

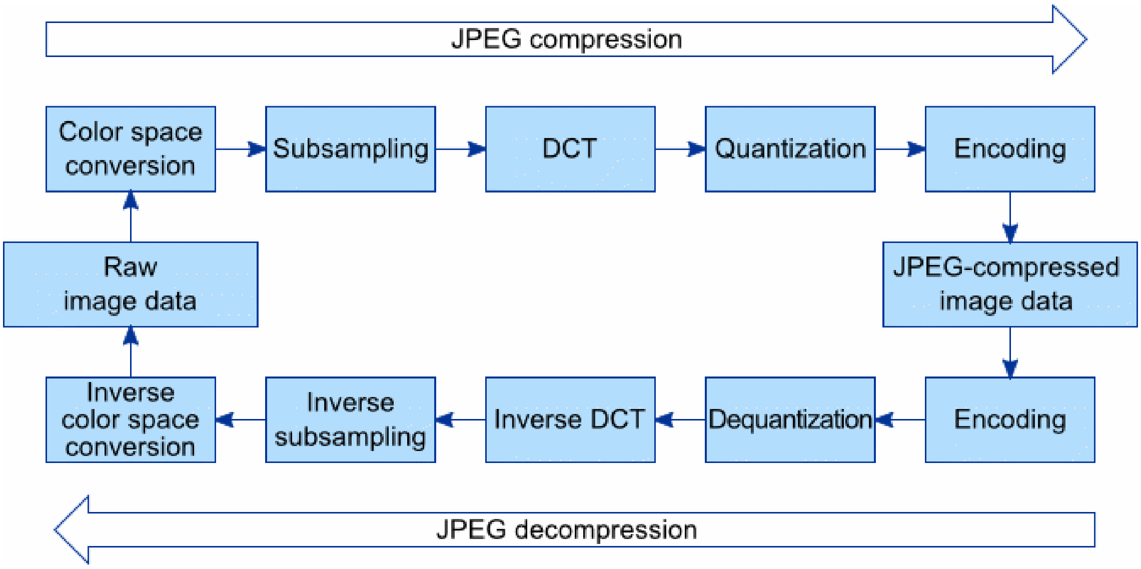


Figure 3.1: JPEG Algorithm schema.

JPEG compression algorithm can be very efficient for certain images but it all depends on the information in the image. There is no ”shortcut” to find the best compression setting and obtaining the best possible quality after the compression. The JPEG algorithm is best suited for photographs and paintings of realistic scenes with smooth variations of tone and color. JPEG is not suited for images with many edges and sharp variations as this can lead to many artifacts in the resultant image. In these situations it is best to use lossless formats such as PNG, TIFF or GIF. It is for this reason that JPEG is not used in medical and scientific applications where the image needs to reproduce the exact data as captured and the slightest of errors may snowball into bigger ones.

If we simplify the JPEG algorithm a little then we can say that JPEG uses 5 basic steps: Color Space Conversion, Subsampling, Block-processing with Discrete Cosine Transformation, Quantization and Variable length encoding.

3.2.1 Colorspace Transformation

The JPEG algorithm allows to encode images that use any type of color space. This is possible since JPEG encodes all components in a color model separately, completely independent of any color space model. The first step of the process is to convert the image from original color space into a different color space called YCbCr. It has three components Y, Cb and Cr: the Y component represents the brightness of a pixel, the Cb and Cr components represent the chrominance (split into blue and red components). This is the same color space as used by digital color television as well as digital video including video DVDs, and is similar to the way color is represented in analog PAL video and MAC but not by analog NTSC, which uses the YIQ color space. The YCbCr color space conversion allows greater compression without a significant effect on perceptual image quality (or greater perceptual image quality for the same compression). The compression is more efficient as the brightness information, which is more important to the eventual perceptual quality of the image. That because the human eyes are much more sensitive to the visual information contained in the high-frequency grey-scale component Y. The other components, Cb Cr are chrominance components, which contain high-frequency color information that our eyes are less sensitive to. This color conversion itself does not lead any data reduction and should be performed for the resulting JPEG file to have maximum compatibility. However, some JPEG implementations in "highest quality" mode do not apply this step and instead keep the colour information in the RGB color model, where the image is stored in separate channels for red, green and blue luminance. This results in less efficient compression, and would not likely be used if file size was an issue.

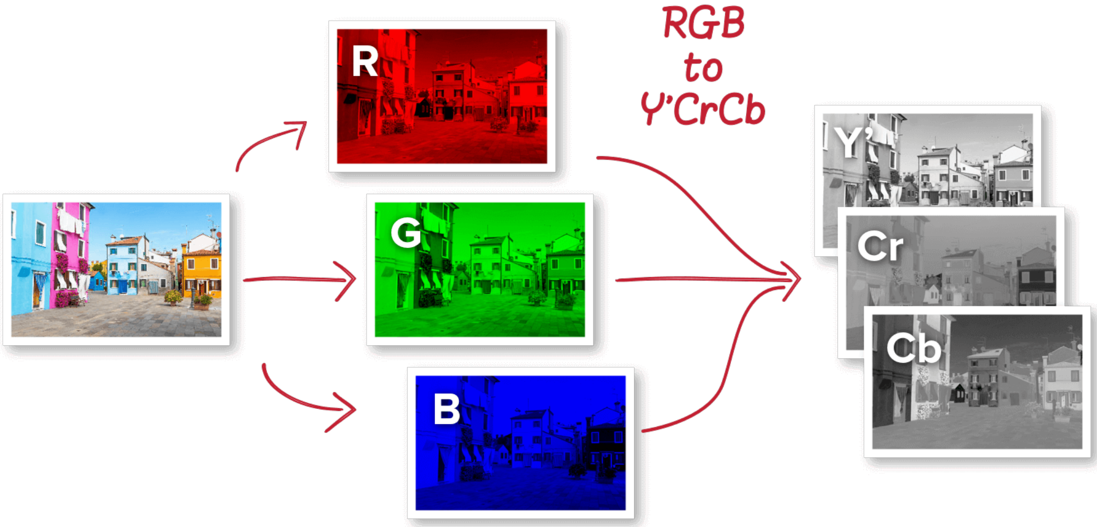


Figure 3.2: Color Conversion from RGB to YCrCb.

3.2.2 Subsampling

This step exploits one of the human eye's weaknesses. Due to the densities of color- and brightness-sensitive receptors in the human eye, humans can see considerably more fine detail in the perceived brightness of an image (the Y component) than in the color of an image (the Cb and Cr components). Using this knowledge is possible to compress images more efficiently removing some of the chrominance information (color) in a signal in favor of luminance data, without significantly affecting picture quality. The transformation into the YCbCr color model enables the next step, which is to reduce the spatial resolution of the Cb and Cr components (called "downsampling" or "chroma subsampling").

This can be achieved by simply using fewer pixels in the chrominance channels. A signal with chroma 4:4:4 has no compression (so it is not subsampled). The first number (in this case 4), refers to the size of the sample. The two following numbers both refer to chroma. They are both relative to the first number and define the horizontal and vertical sampling respectively. The 4:2:2 signal will have half the sampling rate horizontally, but will maintain

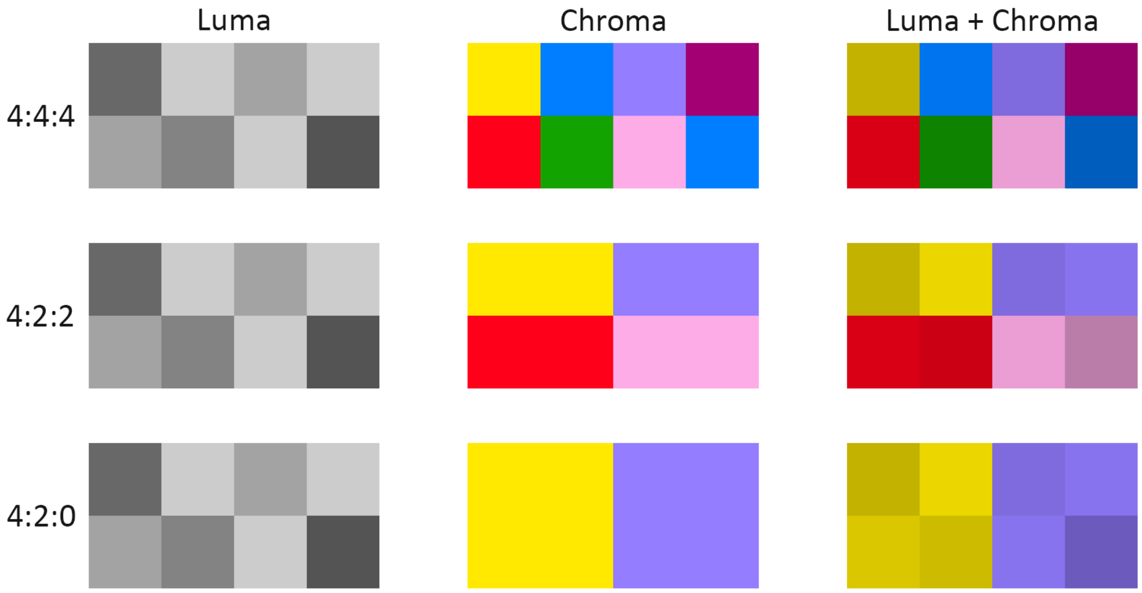


Figure 3.3: Horizontal and Vertical Chroma Subsampling Methods

full sampling vertically. 4:2:0, on the other hand, will only sample colors out of half the pixels on the first row and ignores the second row of the sample completely. The generally used method is the 4:2:0.

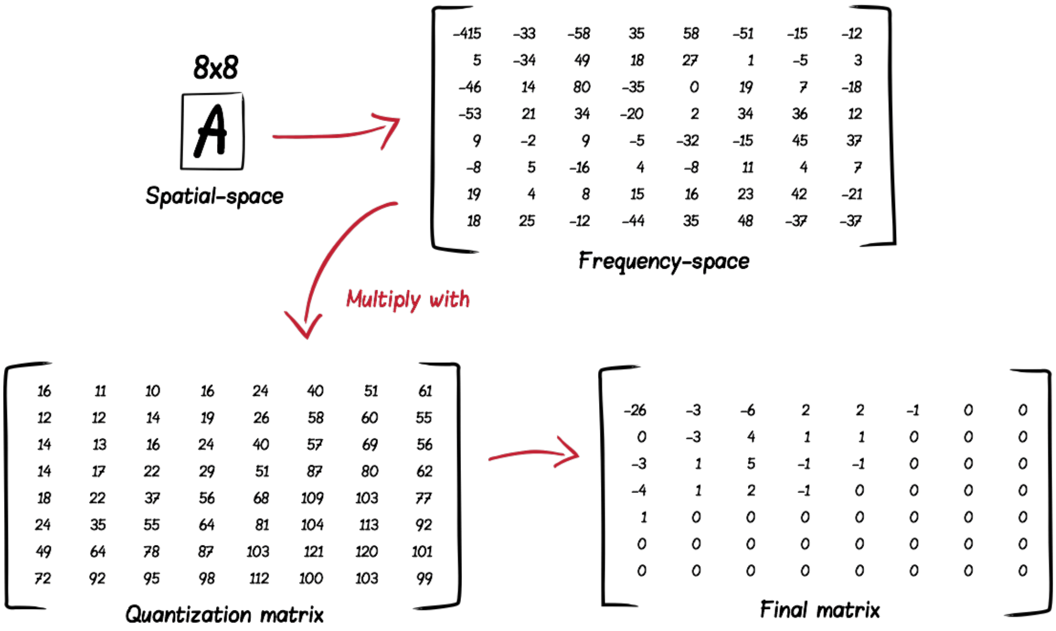
As an example, we imagine an image with the resolution of 1000x1000 pixels. We then transform the image to only use 500x500 pixels for the chrominance component but keep the 1000x1000 pixels for the luminance component. Thus, each chrominance pixel will now cover the same area as a 2x2 luminance block.

3.2.3 Block Splitting and DCT

The block splitting is the process that splice each channel into smaller blocks of 8x8 pixels. Depending on chroma subsampling, this yields (Minimum Coded Unit) MCU blocks of size 8×8 (4:4:4 – no subsampling), 16×8 (4:2:2), or most commonly 16×16 (4:2:0). If the data for a channel does not represent an integer number of blocks then the encoder must fill the remaining area of the incomplete blocks with some form of dummy data. Filling the edge pixels with a fixed color (typically black) creates ringing artifacts along the visible part of the border; repeating the edge pixels is a common technique that reduces the visible border, but it can still create artifacts.

Next these blocks will be transformed into the frequency-space by the Discrete Cosine Transformation. This means that instead of expressing the image in pixels, we transform the block to be expressed in waves. This allows us to separate low and high frequencies. Because of this, we can then apply quantisation, which will allow us to remove high frequencies which the human eye cannot see

This is a very important part of the JPEG compression. There are several versions that can be used of the DCT function but the most common one is DCT-II. When we use DCT, values with the lowest frequency will be clustered around the upper-left corner of the DCT matrix. When performing the DCT transformation, we divide the image into blocks. The standard size of the blocks are 8x8 pixels. This number might look arbitrary, which it actually is. There are however a few reasons for using this block size. First, if the blocks were smaller, the compression would be more troublesome to perform. It would also take longer for the compression to finish. Secondly, if we were to make the blocks bigger then there is a high probability that the image would have larger color gradients between the blocks. Finally, powers of 2 is known to have computational advantages. DCT converts the pixels of the image from the spatial-dimension into the frequency-dimension which makes it possible to separate the low and high frequencies. The reason for separating high and low frequencies lies in the fact that human eyes are unable to see high frequencies. Therefore, removing them have little to no impact of the quality of the image. DCT enables us to separate the high frequencies allowing us to choose which values we want to save. By applying a quantisation matrix on the DCT matrix, we can zero out elements that represent high frequencies, freeing



$$\begin{bmatrix} -415 & -33 & -58 & 35 & 58 & -51 & -15 & -12 \\ 5 & -34 & 49 & 18 & 27 & 1 & -5 & 3 \\ -46 & 14 & 80 & -35 & 0 & 19 & 7 & -18 \\ -53 & 21 & 34 & -20 & 2 & 34 & 36 & 12 \\ 9 & -2 & 9 & -5 & -32 & -15 & 45 & 37 \\ -8 & 5 & -16 & 4 & -8 & 11 & 4 & 7 \\ 19 & 4 & 8 & 15 & 16 & 23 & 42 & -21 \\ 18 & 25 & -12 & -44 & 35 & 48 & -37 & -37 \end{bmatrix}$$

$$\begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

$$\begin{bmatrix} -26 & -3 & -6 & 2 & 2 & -1 & 0 & 0 \\ 0 & -3 & 4 & 1 & 1 & 0 & 0 & 0 \\ -3 & 1 & 5 & -1 & -1 & 0 & 0 & 0 \\ -4 & 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 3.4: Equivalent DCT coefficients, the quantization table and the Quantized DCT coefficients

up memory. Depending on the amount of quantisation we can decide approximately how much we want to compress the image. This is the only step except the subsampling step where the user actually has a chance to influence the end result and impact on how much the image is going to be compressed.

3.2.4 Quantization

The human eye is good at seeing small differences in brightness over a relatively large area, but not so good at distinguishing the exact strength of a high frequency brightness variation. This allows one to greatly reduce the amount of information in the high frequency components. This is done by simply dividing each component in the frequency domain by a constant for that component, and then rounding to the nearest integer. This is the main lossy operation in the whole process. As a result of this, it is typically the case that many of the higher frequency components are rounded to zero, and many of the rest become small positive or negative numbers, which take many fewer bits to store, according to the Quantiozation Tables. Different Tables for Luminance and Chrominance quantization are used, ascertaining tha t the human eye is more sensitive to Luminance than the Chrominance.

3.2.5 Variable Lenght Encoding

Our coefficients will be reduced more and more to zero depending on the quantization in the previous step. It is much more likely that the zero valued coefficients appear in higher frequencies than lower frequencies. The values are sorted with the lower frequencies before the higher frequencies. This will lead to a high probability of having some values at the beginning and then several values with the value of zero. The reason for this is that it is cheaper to store 10x0 rather than 0, 0, 0, 0, 0, 0, 0, 0, 0, 0. The quantized values are scanned in a zig-zag order (starting from top-left corner) reducing so the amount of data significantly. There are two common encoding techniques used for encoding JPEG images, “Huffman” and “Arithmetic encoding”. Both are variable length encoding techniques that are used to map source symbols to a variable number of bits. This is used for compressing and decompressing information, in this case, JPEG images.

The Huffman encoding is a statistical data compression technique that represents the symbols of an alphabet by reducing the average code length. The Huffman code is mconsidered optimal if all the symbols have probability of integral powers of 1/2. A much simplified example of the Huffman encoding would be if we imagine our image as the string ”HHHHHUFFFFFF”. Instead of writing the entire string as it is we could instead write ”Hx5UFx6”. The Huffman encoding achieves this by building a Huffman code tree. It is possible to create many different Huffman codes for a given frequency. However, the total compressed length will always be the same for that given frequency.

Huffman encoding has been and still is the standard encoding process of JPEG images. However, there is actually one encoding technique that can be considered more effective and that is Arithmetic encoding. As seen previously, Huffman encoding is considered optimal when all the symbols probability are integral powers of 1/2. This restriction is not affecting the Arithmetic encoding. What makes Arithmetic encoding different is that it represents a number by an interval of real numbers between one and zero. This means that the interval to represent the message (in this case the representation of the image) becomes smaller and smaller the longer the message gets. If the message contains successive symbols then the interval will be reduced in accordance with the probability of that symbol.This leads to adding fewer bits to the message.

3.3 Effects of JPEG Compression

JPEG compression artifacts blend well into photographs with detailed non-uniform textures, allowing higher compression ratios. Notice how a higher compression ratio first affects the high-frequency textures in the upper-left corner of the image, and how the contrasting lines become more fuzzy. The very high compression ratio severely affects the quality of the image, although the overall colors and image form are still recognizable. However, the precision of colors suffer less (for a human eye) than the precision of contours (based on luminance). This justifies the fact that images should be first transformed in a color model separating the luminance from the chromatic information, before subsampling the chromatic planes (which may also use lower quality quantization) in order to preserve the precision of the luminance plane with more information bits.

Chapter 4

Forensics on Double Compressed JPEG Images

4.1 Related Work

In the last ten years Image Forensics, has developed at a growing rhythm and in particular the interest has been focused on recovering coefficients of the JPEG quantization matrix used to compress an image at the time of shooting (i.e., when the image has been created), when for some reasons this information is no more available in the Exif metadata.

This scenario may include the primary quantization coefficients of an image that has been doubly JPEG compressed, or the retrievial of the compression matrix of an uncompressed image previously JPEG compressed, since in both these cases the values of the primary compression steps are lost. At the state of the art the principal studies are based on DCT coefficients, Benford's Law or Fourier Coefficients, Neural Networks, Factor Histogram and considerations on noise.

In [54] and [55], Fan and de Queiroz describe a method to determine if an image in bitmap format has been previously JPEG-compressed and further to estimate the quantization matrix used. They present a method for the maximum likelihood estimation (MLE) of JPEG quantization steps and showed its reliability via experimental results. It proposes (for the first time in multimedia forensics) an estimation of the lattice structure in the DCT domain, even if limited on the case of quantization performed for a JPEG compressed image without downsampling of color components. As a critical remarks, we point out that both methods are limited to $QF \leq 95$.

In [33], Bianchi et al. started to face with a particular scenario that they called Single Compression Forgery for JPEG images. This is the situation in which a part of a JPEG image is patched over an uncompressed image (copy-paste or cut/past operation) and the result is JPEG compressed. The core of the method is the use of Bayesian inference to assign to each DCT coefficient a probability of being double quantized, giving the possibility to build a probability map that for every part of the image tells if it is original or tampered.

In [32], which is itself a refinement of [31], the authors build a likelihood map to find the regions that have undergone to a double JPEG compression. This method is important because it takes into account two kinds of traces left by tampering in double-compressed JPEG images: aligned and non-aligned, something that was considered, to the best of our knowledge, only few times before [27] [142]. Also in these papers is underlined the difficulty to correctly estimate q1 when it is $\leq q2$.

In [136], mainly dedicated to tampering detection by means of a proper comparison of the different distributions of DCT coefficients between tampered and non tampered regions, Wang et al built up a mathematical model in which the knowledge of q1 is required. They take also into account the truncation and rounding errors, referring to the work of Bianchi et al [31]. We want to point out that also in this work the case q1 ¡ q2 gives unsatisfactory results.

In [59], which refines what has been proposed by the authors in [146], Galvan et al. focus on the determination of first quantization step in double compressed JPEG images, assuming that q1 > q2. The main novelties of the proposed approach, which improves what has been proposed by the authors in [58], are related to the filtering strategy adopted to reduce the amount of noise in the input data (DCT histograms), and on the design of a novel function with a satisfactory q1-localization property. In [128] the authors use a method based on the combination of the quantization effect and the statistics of Discrete Cosine Transform (DCT) coefficient characterized by the statistical model. The analysis of quantization effect is performed within a mathematical framework, which justifies the relation of local maxima of the number of integer quantized forward coefficients with the true quantization step. From the candidate set of the true quantization step given by the previous analysis, the statistical model of DCT coefficients is used to provide the optimal quantization step candidate.

In [56] has presented a novel method to estimate q1 by introducing the mean square error (MSE) sequence of ratios among DCT coefficient histogram bins, Xue et al. formulate the relationship between its periodic fluctuation and q1. And in order to enhance the periodic effect, they propose a strategy to adjust the histogram. Then, based on MSE sequence, several q1 candidates can be obtained. Finally by histogram comparison, the estimated quantization step is selected from the candidates.Only the quality factors (Q1,Q2) in the range of 50 to 90 with step of 5 and the first five AC coefficients in zig-zag order are considered. In [46] the estimation of first quantization value in a non-aligned double JPEG scenario is analyzed. The proposed method investigates the impact of the blocking artifacts induced errors on the DCT histogram and counters these errors via a novel DCT histogram filtering strategy. In addition, the residual noise which is also a recompression artifact is countered utilizing the local rank transform which adaptively filters the residual noise effects under the condition QF2>QF1.

4.1.1 The Galvan Method

In [60] to improve the results of the estimation, a proper filtering strategy together with a function devoted to find the first quantization step, have been designed. Specifically, as shown in Fig. 4.1 the algorithm consists of the following main steps:

- DCT Histogram Filtering. A deep analysis on the con-sequences of both quantization and rounding error has been performed. While indeed the former is well known, the rounding error e manifests itself as peaks spread around the multiples of the quantization step q,as exposed in [134] and has been modeled as an approximate Gaussian noise. Those joint phenomenons will affect the behavior of the second quantization step, thus the magnitude of the DCT coefficients, and consequently its statistics. For those reasons, the filtering strategy must face two kind of noise: the "split noise" and the "residual noise", with the aim to bring the histogram as if the rounding error did not have impact. This module actually provides a set of filtered histograms Hfiltq1i (one for each quantization step q1i in q1min, q1min + 1,...,q1max).

- Proposed Function for Quantization Step Estimation. Once the histogram has been filtered removing (or reducing) the error e, a function exploiting the properties of successive quantizations is needed. Specifically, a function with a q1 localization property is actually valuated over all the histograms Hfiltq1i generating a set of output foutq1i .

- Selection of the Quantization Step Candidates. Starting from the set of output foutq1i, a limited number of first quantization candidates (q1s in Cs) are selected exploiting the q1 localization property of the proposed error function.

- DCT Histogram Based Selection. The double quantization process is then simulated to consider the candidates provided by the other blocks and the best one exploiting directly histogram values is finally selected.

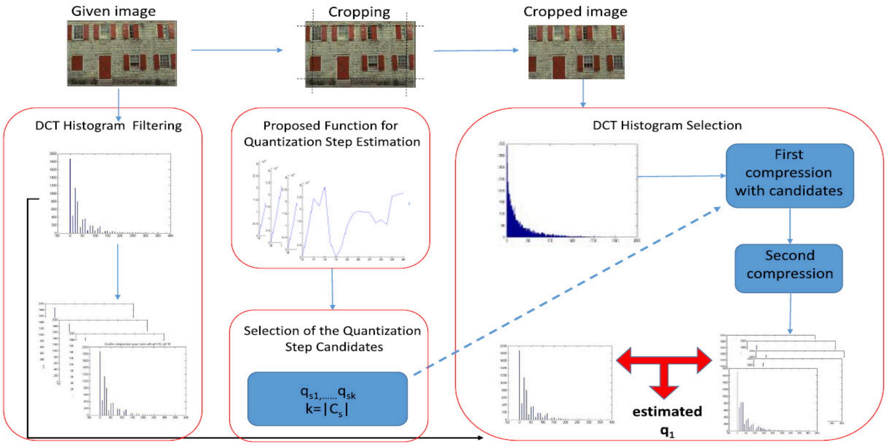


Figure 4.1: Galvan algorithm schema

In [45] Dalmia and Okade highlight that Galvan's method does not account the additional instances of split and residual noise; moreover, the same work does not consider split and residual noise introduced while performing double JPEG compression to form histogram Hq1s leding to incorrect estimation of first quantization value.

4.1.2 Galvan Method Experiment results

Starting from the images, applying JPEG encoding with standard JPEG quantization tables proposed by IJG (Independent JPEG Group), a dataset of double compressed images have been built just considering quality factors (QF1, QF2) in the range 50 to 100 at steps of 10. Taking into account the condition q1 > q2 (i.e., QF2>QF1 in our tests), the final dataset contains 20070 images. Results are then reported with respect to quality and are relative to the first 15 DCT coefficients considered in zig-zag order. This order, used in the standard JPEG, allows sorting the coefficients from the lowest frequency (DC) to the highest frequencies.

To estimate the first quantization matrix, Galvan's technique has been applied only on a minor number of blocks of cropped double compressed image. The analyzed blocks have been chosen in a total random way. By decreasing percentage of considered blocks image, the performances are the same of the original algorithm as shown in fig

The estimation error of the proposed approach is the same of [60], it is close to zero for the DCT coefficients related to low frequency in the DCT domain and does not significantly depend on the specific quality factor employed for the first and second quantization.

To assess the performance of the proposed approach, several tests have been conducted considering double compressed JPEG images, obtained starting from a standard dataset of uncompressed images - Uncompressed Colour Image Database - UCID [123]. Specifically, the UCID (v2) contains 1338 uncompressed TIFF images with a certain variability in terms of scene content(natural, man-made objects, indoor, outdoor, etc.). One of the main features of the UCID dataset is that the images are preserved in their uncompressed state hence suggesting it as a benchmark database for the evaluation of compressed domain image retrieval techniques and allowing to inspect the influence image compression has on the performance of image retrieval algorithms.

4.2 Application of Galvan's Algorithm in Different Scenarios

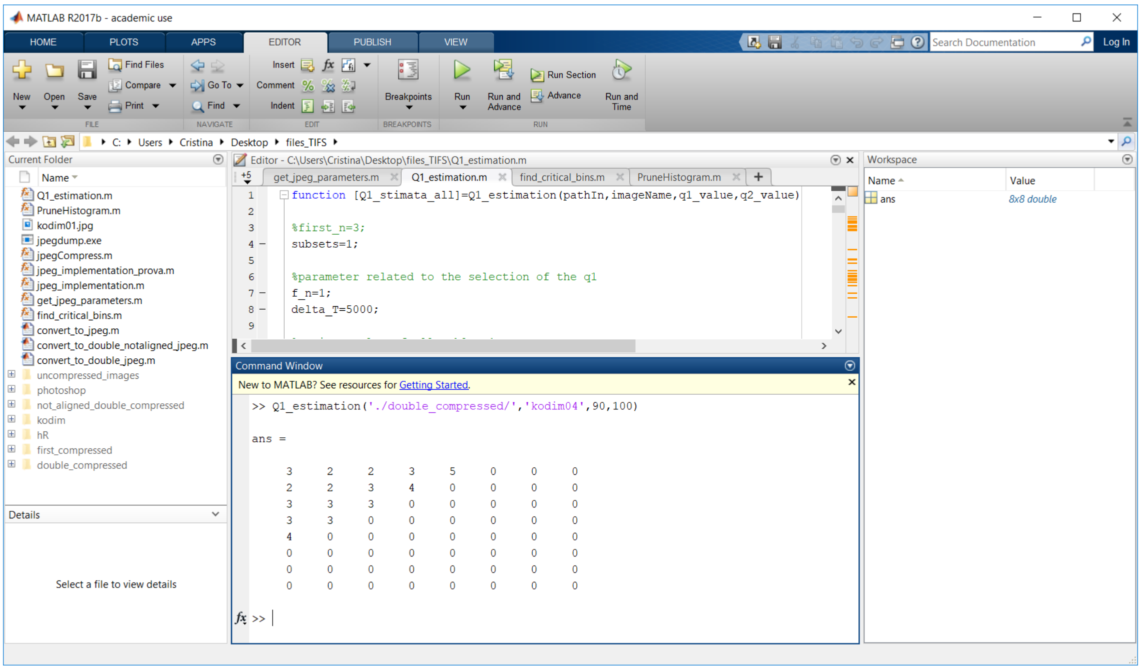
Starting from Galvan's method, given a doubly compressed image, obtained its DCT histogram, a particular error function (4.1) is applied to get a set of first quantization candidates q1.

$$f_e'(y_{i,j}^{(0)}, q_1, q_2, q_3) = \Bigg| \Bigg[\Bigg[\Bigg[\Bigg[\frac{y_{i,j}^{(0)}}{q_1} \Bigg] \times \frac{q_1}{q_2} \Bigg] \times \frac{q_2}{q_3} \Bigg] \times \frac{q_3}{q_2} \Bigg[\times q_2 - \Bigg[\Bigg[\frac{y_{i,j}^{(0)}}{q_1} \Bigg] \times \frac{q_1}{q_2} \Bigg] \times q_2 \Bigg| \qquad (4.1)$$

Considering the DCT coefficients relating to a cropped version of the original image and simulating the double quantization with the previously calculated candidates, it is possible to trace the values of the first quantization matrix by comparing the histograms of the original image with the simulated one. Starting from the same dataset[10] of sample images, the code was reworked to implement first and second compression as the quality factors changed and the relative error percentage was evaluated by applying the same algorithm. Moreover, we tested Galvan's Algorithm in different scenarios:

- High Resolution Image;
- Specific modified parts of the image;
- Not aligned Image;
- Fewer number of blocks.

The implementation and testing phases of the algorithms were carried out entirely using the Matlab programming environment, thanks to the wide range of Toolboxes dedicated to the manipulation and processing of numerical matrices and, more generally, to Image Processing.



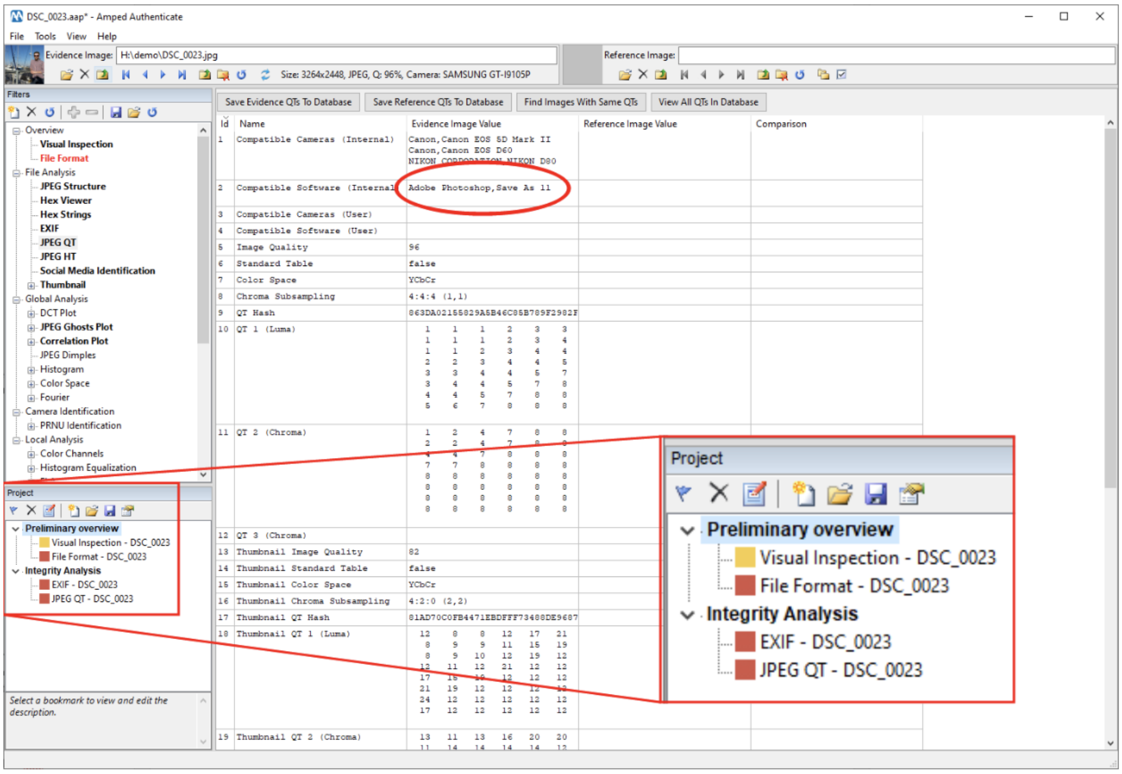
```
function [Q1_stimata_all]=Q1_estimation(pathIn,imageName,q1_value,q2_value)

%first_n=3;
subsets=1;

%parameter related to the selection of the q1
f_n=1;
delta_T=5000;
```

New to MATLAB? See resources for Getting Started.

```
>> Q1_estimation('./double_compressed/','kodim04',90,100)

ans =

     3     2     2     3     5     0     0     0
     2     2     3     4     0     0     0     0
     3     3     0     0     0     0     0     0
     3     3     0     0     0     0     0     0
     4     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0

fx >> 
```

Figure 4.2: An image of Matlab Environment used for the implementation of the code

The results obtained, for some sample images, were compared with the values obtained with the use of Amped Authenticate[2], a software package for authenticating images and detecting tampering on digital photos. Amped Authenticate provides a suite of different tools to determine if an image is an unaltered original, an original generated by a specific device, or the result of manipulation with photo editing software.



Evidence Image: H:\demo\DSC_0023.jpg

Size: 3264x2448, JPEG, Q: 96%, Camera: SAMSUNG GT-I9105P

Id	Name	Evidence Image Value	Reference Image Value	Comparison
1	Compatible Cameras (Internal)	Canon,Canon EOS 5D Mark II Canon,Canon EOS D60 NIKON CORPORATION,NIKON D80		
2	Compatible Software (Internal	Adobe Photoshop,Save As 11		
3	Compatible Cameras (User)			
4	Compatible Software (User)			
5	Image Quality	96		
6	Standard Table	false		
7	Color Space	YCbCr		
8	Chroma Subsampling	4:4:4 (1,1)		
9	QT Hash	863DA02155829A5B46C85B789F2982F		
10	QT 1 (Luma)	1 1 1 2 3 3 1 1 1 2 3 4 1 1 2 3 4 4 2 2 3 4 4 5 3 3 4 4 5 7 3 4 4 5 7 8 4 4 5 7 8 8 5 6 7 8 8 8		
11	QT 2 (Chroma)	1 2 4 7 8 8 2 2 4 7 8 8 4 4 7 8 8 8 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8		
12	QT 3 (Chroma)			
13	Thumbnail Image Quality	82		
14	Thumbnail Standard Table	false		
15	Thumbnail Color Space	YCbCr		
16	Thumbnail Chroma Subsampling	4:2:0 (2,2)		
17	Thumbnail QT Hash	81AD70C0FB4471EBDFFF73488DE9687		
18	Thumbnail QT 1 (Luma)	12 8 8 12 17 21 8 9 9 11 15 19 8 9 10 12 19 12 12 11 12 21 12 12 17 15 10 12 12 12 21 19 12 12 12 12 24 12 12 12 12 12 17 12 12 12 12 12		
19	Thumbnail QT 2 (Chroma)	13 11 13 16 20 20 11 14 14 14 14 12		

Select a bookmark to view and edit the description.

Figure 4.3: An image of the Amped Authenticate software

4.3 Results and Future Works

In this section we present results of our analysis. Unfortunately in some cases they have been inconsistent, in others they have suggested possible future developments.

HIGH RESOLUTION

The error analysis was carried out with high resolution images or manipulated with Photoshop and therefore with different quantization matrices. In particular we then moved on to the study of very high resolution images using the RAISE dataset http://loki.disi.unitn.it/RAISE/. The algorithm works correctly and maintains the same error rates but computational time grows exponentially.

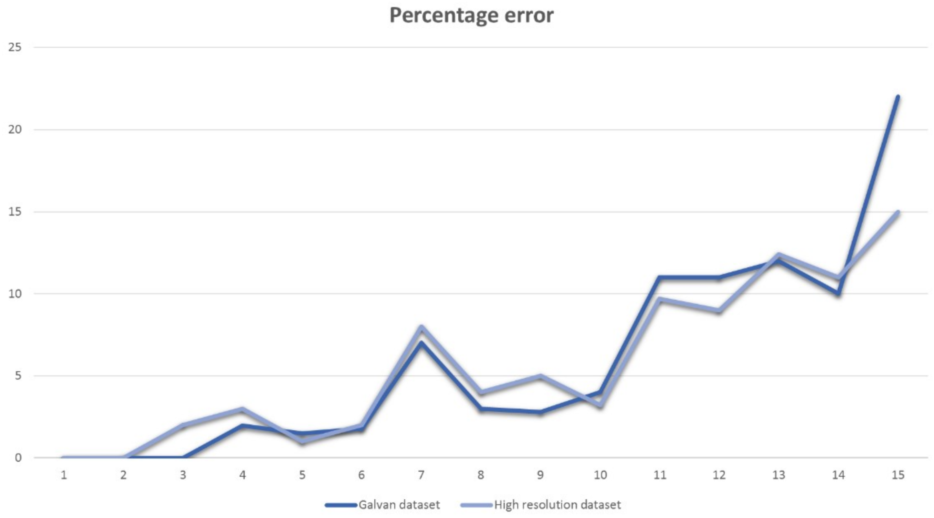


Figure 4.4: Percentage error v/s DCT coefficient position for first quantization value under experimental settings with high resolution images compared with the percetange error of Galvan work

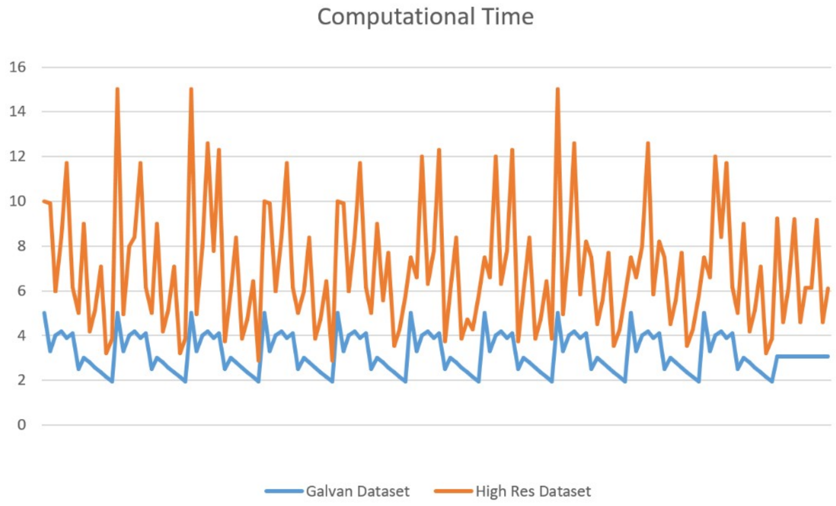


Figure 4.5: Computational time under experimental settings with high resolution images compared with the computational time of Galvan work

IMAGE BLOCKS ANALYSIS

Attention was therefore paid to the behavior of Galvan's algorithm on portions of the image how shown in figure 4.6.

In particular, using a dataset on which a cut and paste operation was performed on a random quadrant of the image between first and second compression. The goal was to verify if the quantization matrix was different only in the quadrant subject to forgery, but the results turned out to be inconsistent.

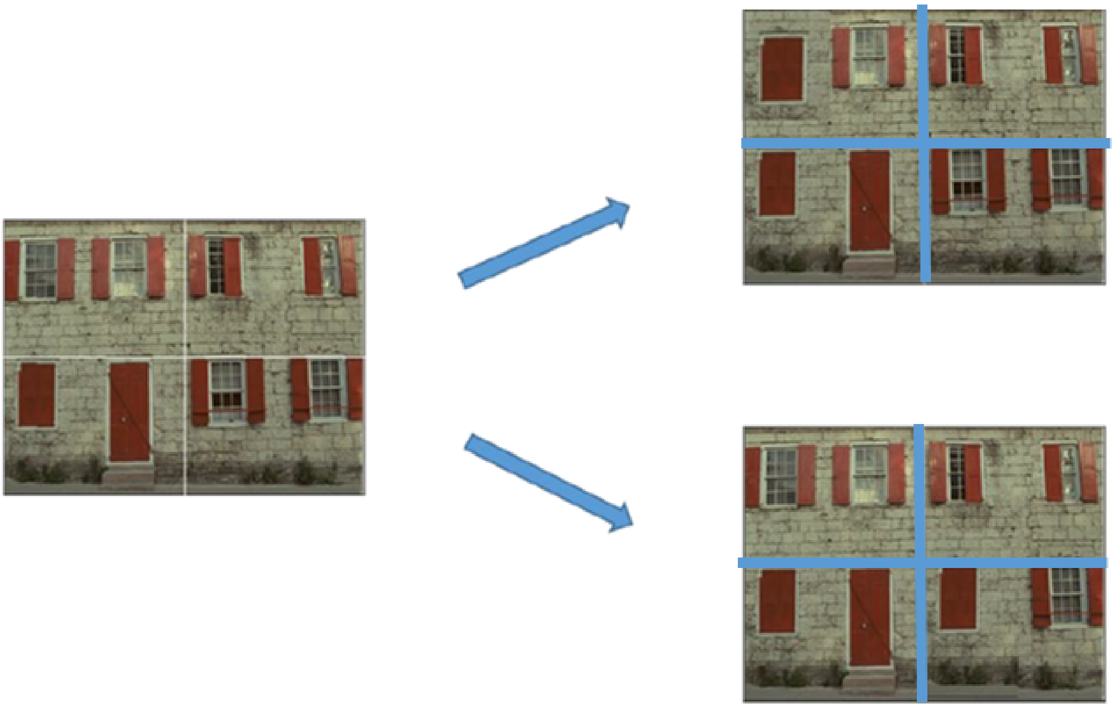
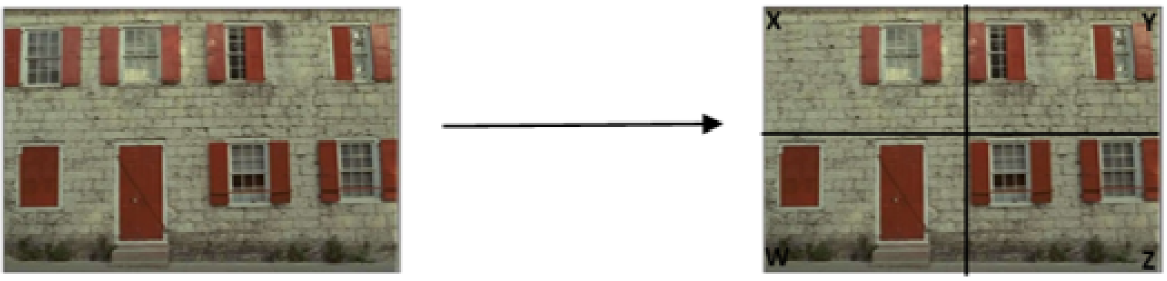


Figure 4.6: An example of image with a cut and paste operation performed on a quadrant of the image between first and second compression



X block

4	11	2	4	4	0	0	0
2	2	14	4	0	0	0	0
	4	4					
7	4						
3							

Y block

13	9	8	13	19	0	0	0
10	10	11	15	0	0	0	0
11	10	13					
11	14						
14							

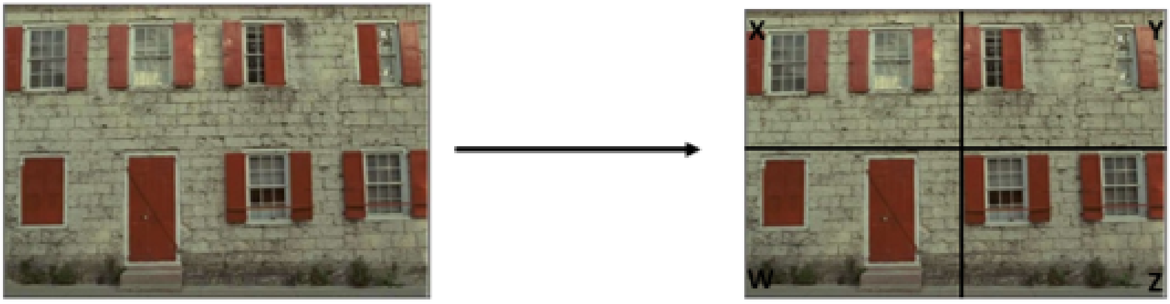
W block

13	9	8	13	19	0	0	0
10	10	11	15	0	0	0	0
11	10	13					
11	14						
14							

Z block

13	9	8	13	19	0	0	0
10	10	11	15	0	0	0	0
11	10	13					
11	14						
14							

Figure 4.7: An image with a cut and paste operation performed on the first quadrant of the image between first and second compression. Only the quantization table of the first block changes.



X block

4	11	2	4	4	0	0	0
2	2	14	4	0	0	0	0
14	4	4					
7	4						
3							

Y block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

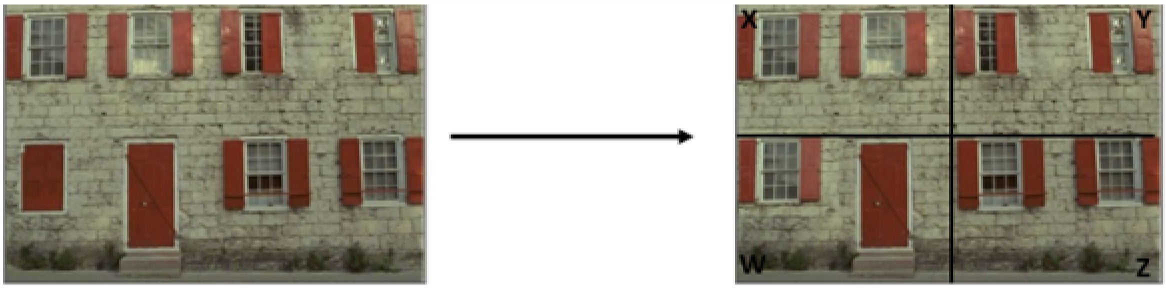
W block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

Z block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

Figure 4.8: An image with a cut and paste operation performed on the second quadrant of the image between first and second compression. Only the quantization table of the first block changes.



X block

4	11	2	4	4	0	0	0
2	2	14	4	0	0	0	0
14	4	4					
7	4						
3							

Y block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

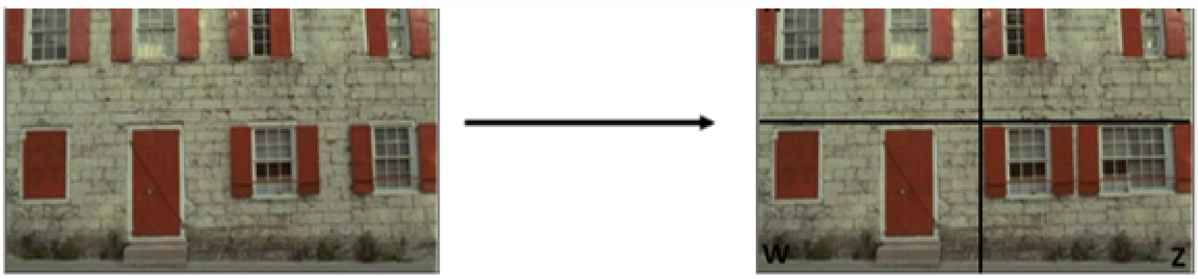
W block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

Z block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

Figure 4.9: An image with a cut and paste operation performed on the third quadrant of the image between first and second compression. Only the quantization table of the first block changes.



X block

4	11	2	4	4	0	0	0
2	2	14	4	0	0	0	0
14	4	4					
7	4						
3							

Y block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

W block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

Z block

16	11	10	16	24	0	0	0
12	12	14	19	0	0	0	0
14	13	16					
14	17						
18							

Figure 4.10: An image with a cut and paste operation performed on the fourth quadrant of the image between first and second compression. Only the quantization table of the first block changes.

NOT-ALIGNED JPEG IMAGES

In this field of research, most of the studies concern aligned double jpeg images, while few concern the more complex case of non-aligned double jpeg images. Aligned double JPEG images are those in which in both compression operations the 8x8 block grids of the JPEG standard line up and overlap. In reality, since these are manipulated images where cropping operations are often used, the most realistic case concerns images in which in the second JPEG compression operation the 8x8 grid does not overlap the first, but has a pixel shift that can go from 1 to 7 pixels.

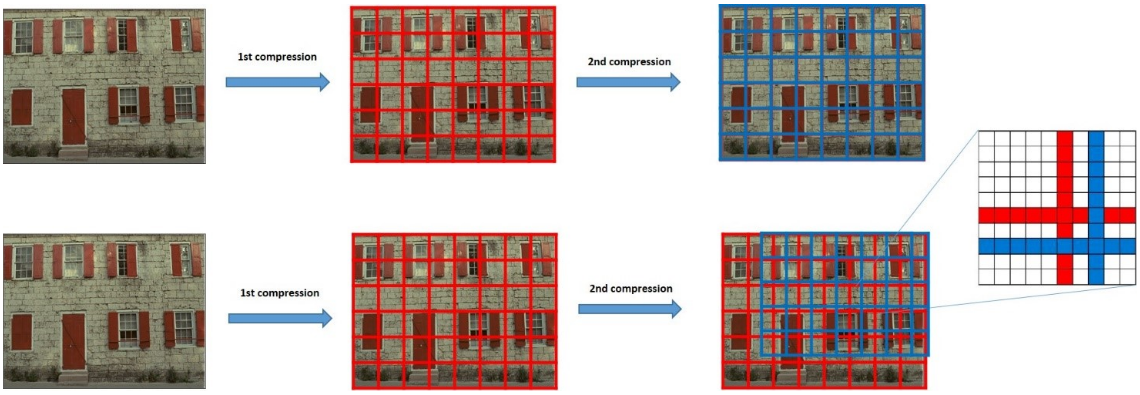


Figure 4.11: Aligned and Not Aligned Double JPEG compressed images.

Since the JPEG standard is based precisely on the processing of 8x8 blocks, in the case of misaligned images the so-called Block Artifact Error is generated, which makes the analysis of the images more complex so that most of the methods present in the literature are not directly usable for non-aligned JPEG images. To apply these algorithms to NADJPEG images it is necessary to find the misalignment shift between the first and second DCT step of the JPEG cycle. The estimation of the degree of misalignment is very useful in forensic applications as it could indicate crop or splicing which serves as a starting point for understanding the NA-DJPEG scenario in detail. One of the methods used is based on the Local Pixel Difference (LPD)[46]. The first step in this proposed technique concerns the estimation of the difference between neighboring pixels, performed for all the pixels of the image using the following formula:

$$LPD(x, y) = I(x, y) + I(x + 1, y + 1) - I(x, y + 1) - I(x + 1, y) \tag{4.2}$$

Once the vertical and horizontal boundaries have been determined, we move on to the estimate of the shift (r, c)[44] of the misaligned image by calculating the following scores:

$$HBS(k) = \sum_{m=1}^{\frac{M}{T}-1} \sum_{j=1}^{N} Horizontal\_Boundary(k + m.T, j) \tag{4.3}$$

$$VBS(k) = \sum_{m=1}^{\frac{N}{T}-1} \sum_{i=1}^{M} Vertical\_Boundary(i, k + m.T) \tag{4.4}$$

where, ”T” is the periodicity which is equal to 8 for standard JPEG, ”k” is a set of integer values between 1 and 7 and M × N is the size of the image under examination. However, this technique has different limits in terms of error and computational times, so we focused on the the study of the error function of the coefficients of the DCT matrix. Particular attention was paid to the presence of peaks in the vicinity of some values. It is hypothesized that the error increases at the points where the image is cropped or is due to the ”zig-zag” order in the reading of the DCT coefficients and this confirmation could be useful for estimating the degree of slippage between the two grids.

FEWER IMAGE BLOCKS ANALYSIS

Finally, it was decided to evaluate the Galvan algorithm by reducing the number of blocks being analyzed. Initially by evaluating a cropped part of the image and then a random % of blocks. The study highlighted that in the image datasets used a reduction of up to a maximum of 30% in the number of blocks did not lead to an increase in the error. Starting

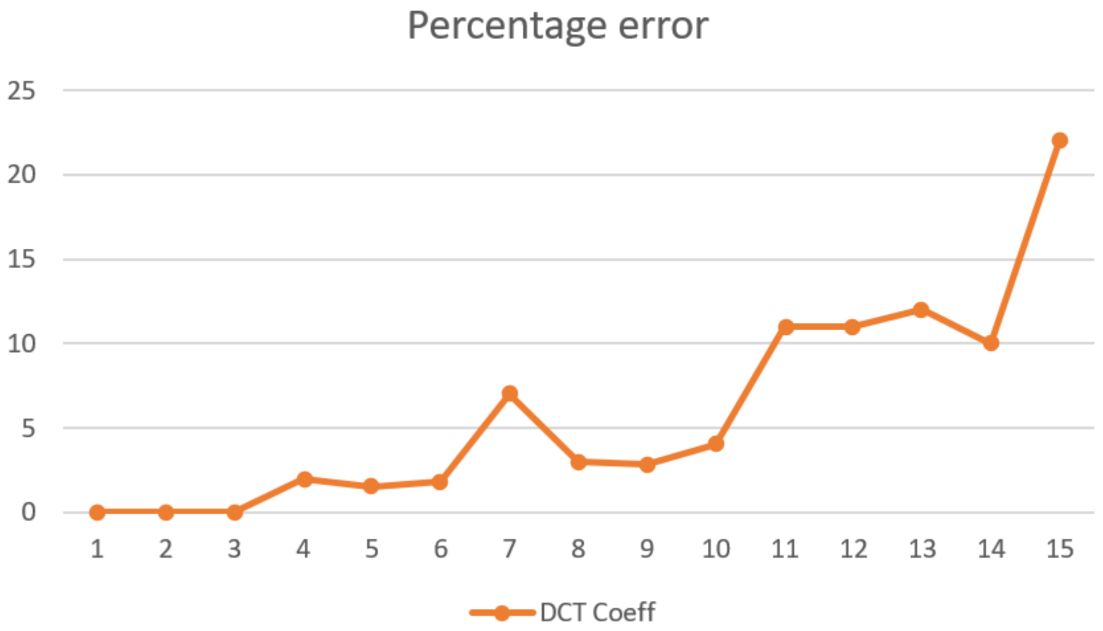


Figure 4.12: Percentage error v/s DCT coefficient position for first quantization value at fixed quality factor.

from this result we thought of creating a method to select which blocks to insert and which to leave out. There have been many roads traveled and attempts made, many of these have not brought appreciable results in terms of error, others have proved to be very expensive from a computational point of view (such as the greedy cluster approach), while it should be noted that the selection of blocks with an average DC coefficient provided results worthy of further study (without however going down to less than 50% of the blocks). Therefore it is believed that we are faced with a law that regulates this very complex problem, so attempting an approach based on deep learning techniques and on the latest generation of neural networks may represent the best choice to continue with the research activity.

Chapter 5

Forensics Analysis on Deep Fake Images

5.1 Artificial Intelligence in Digital Media: Deepfakes

One of the most terrifying phenomenon nowadays is the Deepfake: the possibility to automatically replace a person's face in images and videos by exploiting algorithms based on deep learning. This paper will present a brief overview of technologies able to produce Deepfake images of faces. A forensics analysis of those images with standard methods will be presented: not surprisingly state of the art techniques are not completely able to detect the fakeness. To solve this, a preliminary idea on how to fight Deepfake images of faces will be presented by analysing anomalies in the frequency domain.

Artificial intelligence technologies [114] are evolving so rapidly that unthinkable new applications and services have emerged: one of them is the DeepFake. DeepFakes refers to all those multimedia contents synthetically altered or created by exploiting machine learning generative models. DeepFakes are image, audio or video contents that appear extremely realistic to humans specifically when they are used to generate and/or alter/swap image of faces. Various examples of DeepFake, involving celebrities, have already be seen on the internet: the insertion of Nicholas Cage[1] in movies where he did not act like "Fight Club" and "The Matrix" or the impressive video in which Jim Carrey[2] plays Shining in place of Jack Nicholson. Other more worrying examples are the video of Obama (Figure 5.1(a)), created by Buzzfeed[3] in collaboration with Monkeypaw Studios, or the video in which Mark Zuckerberg[4] (Figure 5.1(b)) claims a series of statements about the platform's ability to steal its users' data.

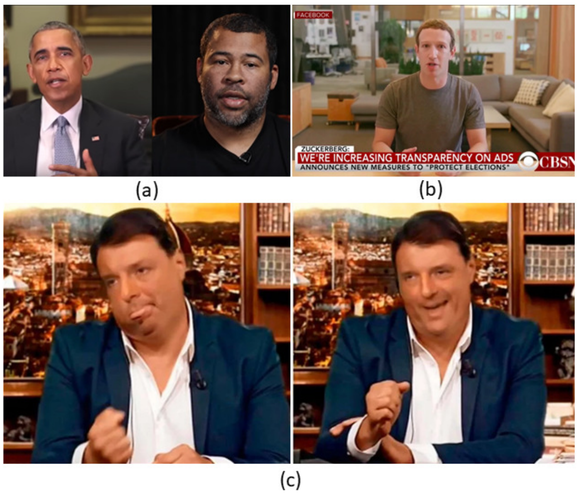
These Deepfakes have already been spread by mass media, also in Italy, where the satirical

[1] https://www.youtube.com/watch?v=-yQxsIWO2ic

[2] https://www.youtube.com/watch?v=Dx59bskG8dc

[3] https://www.youtube.com/watch?v=cQ54GDm1eL0

[4] https://www.youtube.com/watch?v=NbedWhzx1rs


FACEBOOK
ZUCKERBERG:
WE'RE INCREASING TRANSPARENCY ON ADS
ANNOUNCES NEW MEASURES TO "PROTECT ELECTIONS"
CBS

(a) (b)

(c)

Figure 5.1: Several examples of DeepFake: (a) Obama, created by Buzzfeed in collaboration with Monkeypaw Studios; (b) Mark Zuckerberg, created by artists Bill Posters and Daniel Howe in partnership with advertising company Canny; (c) Matteo Renzi, created by "Striscia la Notizia".

tv program "Striscia La Notizia" [5], broadcasted in September 2019 a video of the ex-premier Matteo Renzi talking about his colleagues in a "not so respectful" way (Figure 5.1 (c)). As we can imagine, DeepFakes may have serious repercussions on the veracity of the news spread by the mass media while representing a new threat for politics, companies and personal privacy.

Deepfakes are evolving quickly and are becoming dangerous, not just for the reputation of the victims but also for security. In this dangerous scenario, tools are needed to unmask the deepfakes detect them or, at least, to mitigate the potential harm and abuse that can be done be means of these multimedia contents.

Several big companies, from Facebook to Microsoft, have decided to take action against this phenomenon: Google has created a database of fake videos [119] to support researchers who are developing new methods to detect them while Facebook and Microsoft have launched

[5] https://www.striscialanotizia.mediaset.it/video/

the Deepfake Detection Challenge initiative[6] which invites people from all over the world to create new tools to detect deepfakes and manipulated media.

The motivation of this work is straightforward, after a brief overview of the state of the art in order to better understand what the technologies able to produce Deepfake are, a preliminary forensics analysis will be carried out. A first contribution to the field will be demonstrating that it is possible for the image forensics expert to find anomalies that could be related to how the Deepfake image is made. In fact standard image forensics tools are able to highlight some anomalies which the expert can analyze in deep to find specific anomalies in the frequency domain. In particular the anomalies, after Fourier transform, will be shown and will make clear that each kind of DeepFake creation technology has an easily detectable pattern. These evidence - as a preliminary result to the field - could lead to further and more sophisticated (even automatic) detection and analysis techniques.

The remaining part of this chapter is organized as follow. Section 5.2 presents an overview of DeepFake creation technologies. Section 5.3 investigates how Image Forensics can fights DeepFakes. The proposed method is described in Section 5.4.

[6] https://deepfakedetectionchallenge.ai/

5.2 An overview on Generative Technologies

5.2.1 From Generative Model to GAN

A generative model describes how data is generated, in terms of a probabilistic model.

In the scenario of supervised learning[104], a generative model estimates the joint probability distribution of data P(X, Y) between the observed data X and corresponding labels Y.

Examples of popular generative models are:

- Gaussian mixture model
- Hidden Markov model
- Probabilistic context-free grammar
- Bayesian network
- Averaged one-dependence estimators
- Latent Dirichlet allocation
- Boltzmann machine
- Variational autoencoder
- Generative adversarial network
- Flow-based generative model

Unlike the generative modelling, which studies from the joint probability P(X, Y), the discriminative modeling [75][103] studies the P(Y|X) or the direct maps the given unobserved variable (target) x a class label y depended on the observed variables (training samples). Standard examples of this type of discriminative classifiers are:

- k-nearest neighbors algorithm
- Logistic regression
- Support Vector Machines
- Maximum-entropy Markov models
- Conditional random fields
- Neural networks

5.2.2 Machine Traslation

The possibility of using digital machines to translate documents starting from a natural human languages was introduced for the first time in the second half of the decade of the 1940s by Warren Weaver [15], considered a pioneer in this field. The memorandum entitled simply "Translation" written by Weaver in July 1949 at Carlsbad, New Mexico, was probably the most influential publication in Machine Translation (MT) and . Weaver's memorandum suggests more advantageous methods than any simple word for word approach and was the direct stimulus for the beginnings of research throughout the world. This field had a rapid development and since 2013 all research paper was based on the traditional phrase-based statistical machine translation [112][113][109][38][98]. A new approach for statistical machine translation model, inspired by the trend of deep representational learning has been proposed in 2013 by Kalchbrenner and Blunsom [102]. This new approach is based purely on neural network and conduct the automatic translation workflow very differently with the tradional phrase-based statistical machine translation methods. The Neural Machine Translation (NMT) model uses the Neural Network to learn the model and maximize the translation performance using only a fraction of the memory needed by traditional phrase-based statistical machine translation models. The NMT model consist mainly in two recurrent neural network steps of encoder and decoder [81][72]. The encoder extracts a fixed-length vector representation from a variable-length input sentence and the decoder generates a correct, variable length target translation. The early models have been greatly improved through the use of the novel type of neural network as in [82] where the authors presents a new neural machine translation model based on a gated recursive convolutional neural network. Nowadays this encoder-decoder architecture represents the dominant approach in Machine Translation issues. Nevertheless it suffers from the constraint that all input sequences are forced to be encoded to a fixed length internal vector. This can limit the performance of these networks in the case we consider a very long sentences to translate. The solution of this problem is the use of an attention mechanism [50] that allows to training the model to learn to pay selective attention on the input sequence of text and relate them to each word of the output sequence that is decoding. The encoder-decoder recurrent neural network architecture with attention [140] is currently the state of the art on some benchmark problems for machine translation and it is used in the heart of the Google Neural Machine Translation System. To improve the robustness of the neural machine translation models has been recently used also the generative adversarial network to train the NMT model generating sentences which are hard to be discriminated from human translations [140][152]. In this last case is builded a conditional sequence generative adversarial network where are jointly trained two sub adversarial models: a generator of the target language sentence based on the input source-language sentence and a discriminator, conditioned on the source language sentence, predicts the probability of the target language sentence being a human-generated one.

5.2.3 Generative Adversarial Networks

Synthetic audiovisual media can be generated with a variety of techniques. An overview on Media forensics with particular focus on Deepfakes has been recently proposed in[132, 131].

Currently, the most popular of these techniques is the Generative adversarial networks (GANs)due to its flexible applications and realistic outputs. The GANs are a deep neural net architectures formed by two nets, pitting one against the other (thus the "adversarial"). Before going into the details, let's give a quick overview of what GANs are made for and what maths modelling are behind these models. Generative Adversarial Networks belong to the set of generative models. It means that they are able to generate new content.

Generative adversarial networks (GANs) were firstly introduced by Ian Goodfellow [65] and other researchers at the University of Montreal in 2014. They propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D able to estimate the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a min-max two-player game. In the proposed adversarial nets framework, the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution. In its original version it is composed of two components: a generative model, or generator G, and a discriminative model, or discriminator D, both realized through neural networks. The purpose of the generative model is to produce new data, while the discriminative model learns how to distinguish real data from artificially generated ones.

In particular, given a latent space z and a prior distribution $\mathrm{p_z(z)}$, the generator represents a differentiable function $\mathrm{G(z,\Theta_g)}$ that outputs the new data according to a certain distribution pg, where Θ_g are the parameters of the generative model. The discriminator represents as a differentiable D(x; Θ_d), where Θ_d are the parameters of the discriminative model, which outputs the probability that x comes from the distribution of training data p_{data}. The aim is to get a generator that is a good estimator of p_{data}. When this happens, the discriminator is "deceived" and do not longer distinguish if the samples come from p_{data} or p_g. So both networks work in a competitive training. The discriminative network[2] is trained to maximize the probability of correctly classifying the samples from the training data and the samples generated. At the same time, the generative network is trained by minimizing

$$log(1 - D(G(z)) \tag{5.1}$$

and thus maximizing the probability of the discriminator to consider the samples produced by the generative network coming from p_{data}.

$$MIN_G MAX_D V(D,G) = \mathbb{E}_{x \backsim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \backsim p_z(z)}[log(1 - D(G(z)))] \tag{5.2}$$

The two networks are alternately trained through error back-propagation, keeping the parameters of the generative model unchanged during discriminator training and, vice versa, keeping the parameters of the discriminative network unchanged during generator training.

In the case of Deepfakes, the G can be thought as a team of counterfeiters trying to produce fake currency, while the D stands to the police, trying to detect the malicious activity. G and D can be made by any kind of generative model, in the original version they are implemented through deep neural networks.

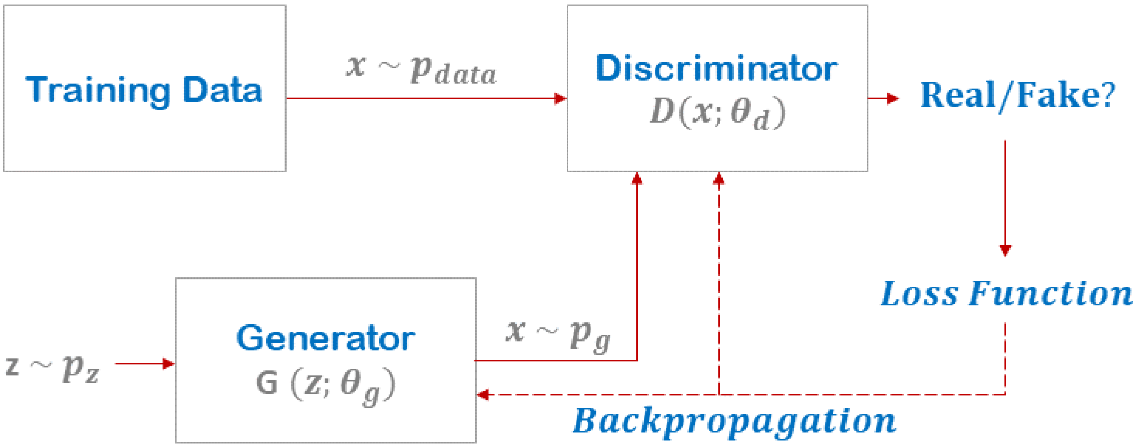


Figure 5.2: Simple schematic view of a Generative Adversarial Network (GAN): the purpose of the Generator is to produce new data, while the Discriminator learns how to distinguish real data from artificially generated one. The two networks are alternately trained through error back-propagation, keeping the parameters of the generative model unchanged during discriminator training and, vice versa, keeping the parameters of the discriminative network unchanged during generator training.

Generating Elements of a Given Distribution

The GAN are strictly connected with the process to generating random variables from a given distribution.

There exist different techniques to generate complex random variables, for example rejecting sampling, Metropolis Hasting algorithm, inverse transform method and others. All these method rely on different mathematical tricks that mainly consist in representing the random variable as the result of an operation or a process.

Between these the inverse transform method is the one that represent the initial step to realize a GAN. It is based on the representation of our complex random variable as the result of a function applied to a uniform random variable that we know how to generate. Considering now a one dimensional example. Let X be a complex random variable we want to sample from and U be a uniform random variable over [0,1] we know how to sample from. A random variable is fully defined by its Cumulative Distribution Function (CDF), a function from the domain of definition of the random variable to the interval [0,1] and defined such that

$$CDF_X = \mathbb{P}(X \leq x) \tag{5.3}$$

Our uniform random variable U is defined such that

$$CDF_U = \mathbb{P}(U \leq u) = u \tag{5.4}$$

The inverse function of the function CDF_X is denoted

$$CDF_x^{-1}(U) \tag{5.5}$$

So, if we define

$$Y = CDF_x^{-1}(U) \tag{5.6}$$

we obtain

$$CDF_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(CDF_x^{-1}(U) \leq y) = \mathbb{P}(U \leq CDF_X(y)) = CDF_X(y) \tag{5.7}$$

So, by defining Y as a function of a uniform random variable we have managed to define a random variable with the targeted distribution.

5.2.4 GAN Taxonomy

There are many types of architecture-variants proposed in the literature [19],[68],[73],[126].

Architecture variant GANs are mainly proposed for the purpose of different applications e.g., image to image transfer [73], image super resolution [36], image completion [121], and text-to-image generation [122]. In [138] Wang at al. provide a taxonomy and an insight on the footprint that current GANs research focuses on the performance improvement. In this

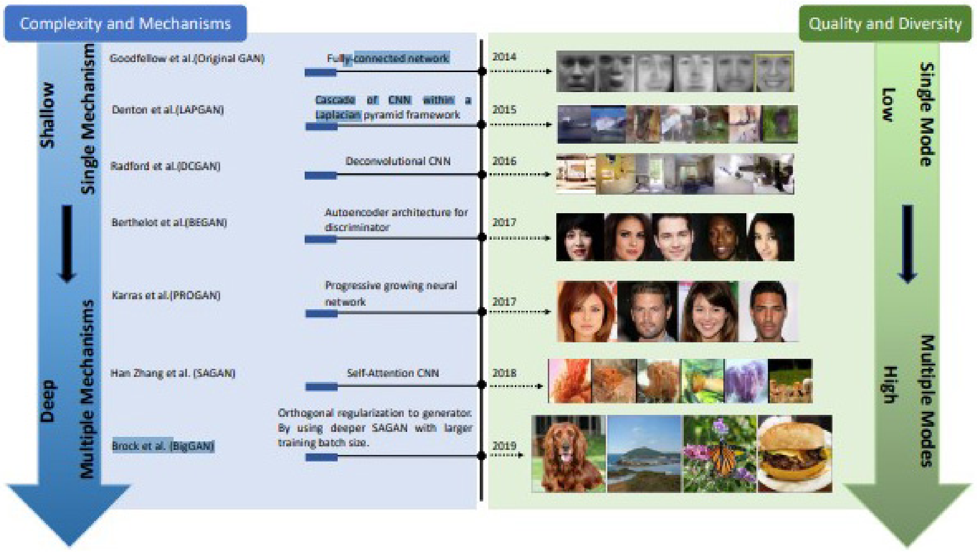


Figure 5.3: Timeline of architecture-variant GANs. Complexity in blue stream refers to size of the architecture and computational cost such as batch size. Mechanisms refer to number of types of models used in the architecture (e.g., BEGAN uses autoencoder architecture for discriminator while a deconvolutional neural network is used for generator. In this case, two mechanisms are used).

section is presented an overview on architecture-variants that helps improve the performance for GANs from the aspects mentioned before.

Fully-connected GAN (FCGAN). The original GAN paper [70] uses fully-connected neural networks for both generator and discriminator. This architecture-variant is applied for some simple image datasets i.e., MNIST [143], CIFAR-10 [18] and Toronto Face Dataset. It does not demonstrate good generalization performance for more complex image types.

Laplacian Pyramid of Adversarial Networks (LAPGAN). LAPGAN is proposed for the production of higher resolution images from lower resolution input GAN [52]. Figure. 3 demonstrates the up-sampling process of generator in LAPGAN from right to left. LAPGAN utilizes a cascade of CNNs within a Laplacian pyramid framework [110] to generate high quality images.

Deep Convolutional GAN (DCGAN). DCGAN is the first work that applied a deconvolutional neural networks architecture for G. Deconvolution is proposed to visualize the features for a CNN and has shown good performance for CNNs visualization[92]. DCGAN deploys the spatial up-sampling ability of the deconvolution operation for G, which enables the generation of higher resolution images using GANs.

Boundary Equilibrium GAN. BEGAN uses an autoencoder architecture for the discriminator which was first proposed in BEGAN [74]. Compared to traditional optimization, the BEGAN matches the autoencoder loss distributions using a loss derived from the Wasserstein distance instead of matching data distributions directly. This modification helps G to generate easyto-reconstruct data for the autoencoder at the beginning because the generated data is close to 0 and the real data distribution has not been learned accurately yet, which prevents D easily winning G at the early training stage.

Progressive GAN (PROGAN). It involves progressive steps toward the expansion of the network architecture [126]. This architecture uses the idea of progressive neural networks first proposed in[16]. This technology does not suffer from forgetting and can leverage prior knowledge via lateral connections to previously learned features. Consequently it is widely applied for learning complex task sequences. Figure. 6 demonstrates the training process for PROGAN. Training starts with low resolution 4 x 4 pixels image. Both G and D start to grow with the training progressing. Importantly, all variables remain trainable throughout this growing process. This progressive training strategy enables substantially more stable learning for both networks. By increasing the resolution little by little, the networks are continuously asked a much simpler question comparing to the end goal of discovering a mapping from latent vectors. All current state-of-the-art GANs employ this type of training strategy and it has resulted in impressive, plausible images, [17].

BigGAN. It [17] has also achieved state-of-the-art performance on the ImageNet datasets.

Its design is based on SAGAN and it has been demonstrated that the increase in batch size and the model complexity can dramatically improve GANs performance with respect to complex image datasets.

There are many types of models proposed in the literature [78, 115, 148, 154]. Architecture variant GANs are mainly proposed for the purpose of different applications e.g., image completion [71], image super resolution [86], text-to-image generation [117] and image to image transfer [154]. The original GAN paper [65] employed fully-connected neural networks for both generator and discriminator. *Laplacian Pyramid of Adversarial Networks* is proposed for the production of higher resolution images from lower resolution input GAN [47]. *Deep Convolutional GAN* is the first work where a deconvolutional neural networks architecture [147] is applied. *Boundary Equilibrium GAN* uses an autoencoder architecture for the discriminator which was first proposed in [151]. *Progressive GAN* involves progressive steps toward the expansion of the network architecture [78]. This architecture uses the idea of progressive neural networks first proposed in [120]. *BigGAN* [35] has also achieved state-of-the-art performance on the ImageNet datasets.

GAN Use Cases

Text to Image Generation.

Scott Reed et others [122] focus their attention in translating text in the form of single-sentence human-written descriptions directly into image pixels. They develop a novel deep architecture and GAN formulation to effectively bridge these advances in text and image modeling, translating visual concepts from characters to pixels and demonstrate the capability of our model to generate plausible images of birds and flowers from detailed text descriptions.

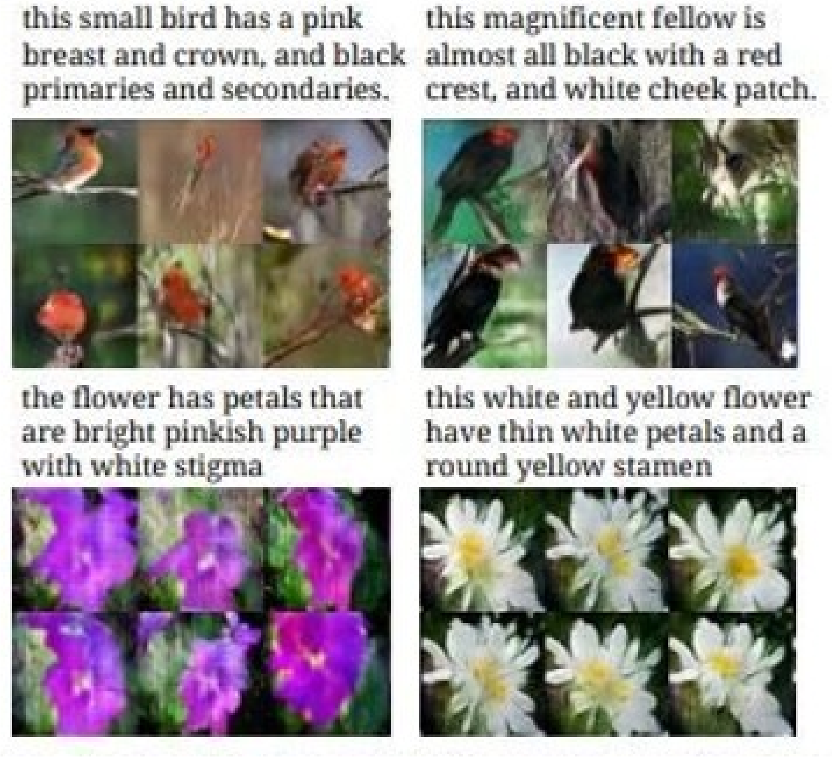

this small bird has a pink breast and crown, and black primaries and secondaries.
this magnificent fellow is almost all black with a red crest, and white cheek patch.
the flower has petals that are bright pinkish purple with white stigma
this white and yellow flower have thin white petals and a round yellow stamen

Figure 5.4: Examples of generated images from text description

This conditional multi-modality is thus a very natural application for generative adversarial networks [70], in which the generator network is optimized to fool the adversarial trained discriminator into predicting that synthetic images are real.

Image to Image Translation. In [53] it investigates conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. They demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks.

Many problems in image processing, computer graphics, and computer vision can be

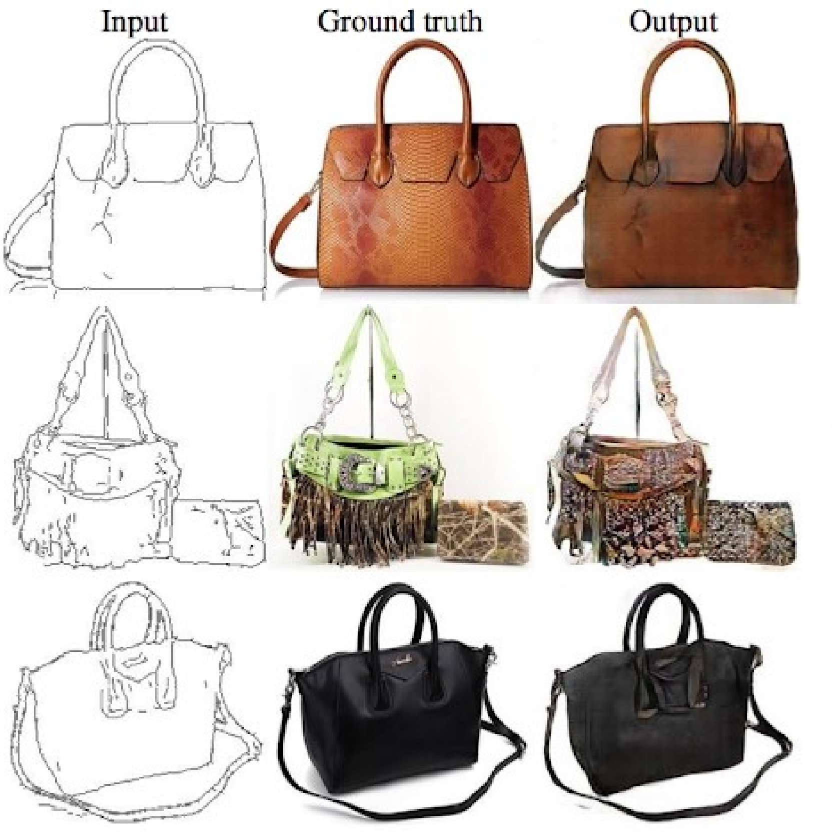


Figure 5.5: Examples of generated images traslation

posed as “translating” an input image into a corresponding output image. Just as a concept may be expressed in either English or French, a scene may be rendered as an RGB image, a gradient field, an edge map, a semantic label map, etc. In analogy to automatic language translation, we define automatic image-to-image translation as the task of translating one possible representation of a scene into another, given sufficient training data.

In this paper, GANs are explored in the conditional setting. Just as GANs learn a generative model of data, conditional GANs (cGANs) learn a conditional generative model [70]. This makes cGANs suitable for image-to-image translation tasks, where we condition on an input image and generate a corresponding output image. Prior and concurrent works have conditioned GANs on discrete labels [99], [51], [61], text, and, indeed, images.

The method also differs from the prior works in several architectural choices for the generator and discriminator. Unlike past work, for this generator it uses a "U-Net"-based architecture [108], and for discriminator a convolutional "PatchGAN" classifier is used.

To optimize networks, they follow the standard approach from [139]: they alternate between one gradient descent step on D, then one step on G. As suggested in the original GAN paper, rather than training G to minimize

$$log(1 - D(G(z)) \tag{5.8}$$

instead train to maximize log D(x,G(x,z)). In addition, they divide the objective by 2 while optimizing D, which slows down the rate at which D learns relative to G.

Increasing Image Resolution.

In the context of super-resolution image [41], the not yet solved problem is how the finer texture details can be recover when it super-resolves at large upscaling factors. The behavior of optimization-based super-resolution methods is principally driven by the choice of the objective function. Recent work has largely focused on minimizing the mean squared reconstruction error. The resulting estimates have high peak signal-to-noise ratios, but they are often lacking high-frequency details. Leding et others in [41] present SRGAN, a generative adversarial network (GAN) for image super-resolution (SR). It is the first framework capable of inferring photo-realistic natural images for 4x upscaling factors. To achieve this, they propose a perceptual loss function which consists of an adversarial loss and a content loss. The adversarial loss pushes their solution to the natural image manifold using a discriminator network that is trained to differentiate between the super-resolved images and original photo-realistic images.

Predicting Next Video Frame.

The ability to predict future states of the environment is the focus of AI. At its core, effective prediction requires an internal model of the world and an understanding of the rules by which the world changes. In [139], it explores the internal models developed by deep neural networks trained using a loss based on predicting future frames in synthetic video sequences, using a CNN-LSTM-deCNN framework. Using a weighted mean-squared error and adversarial loss [70], the same architecture successfully extrapolates out-of-the-plane rotations of computer-generated faces. Furthermore, despite being trained end-to-end to predict only pixel-level information, this Predictive Generative Networks learn a representation of the latent structure of the underlying three-dimensional objects themselves. Importantly, they find that this representation is naturally tolerant to object transformations, and generalizes well to new tasks, such as classification of static images.

5.2.5 Technologies for image creation

In Figure 5.6 is shown a generic GAN architecture used to identify real or fake faces.

Focusing on DeepFake images of face, in [83] Lample et al. an “encoder-decoder” architecture, (**Fader Networks**), able to generate different realistic versions of an input image by varying the values of the attributes was introduces: given an input image x with its attributes y, the encoder maps x to a latent representation z, and the decoder is trained to reconstruct x given (z, y). At inference time, a test image is encoded in the latent space and the user chooses the values of the attributes y that are sent to the decoder. A classifier learns how to predict the y attributes given the latent representation z during training. The encoder-decoder is trained so that the latent representation z must contain enough sufficients information to allow input’s reconstruction while the latent representation must prevent the

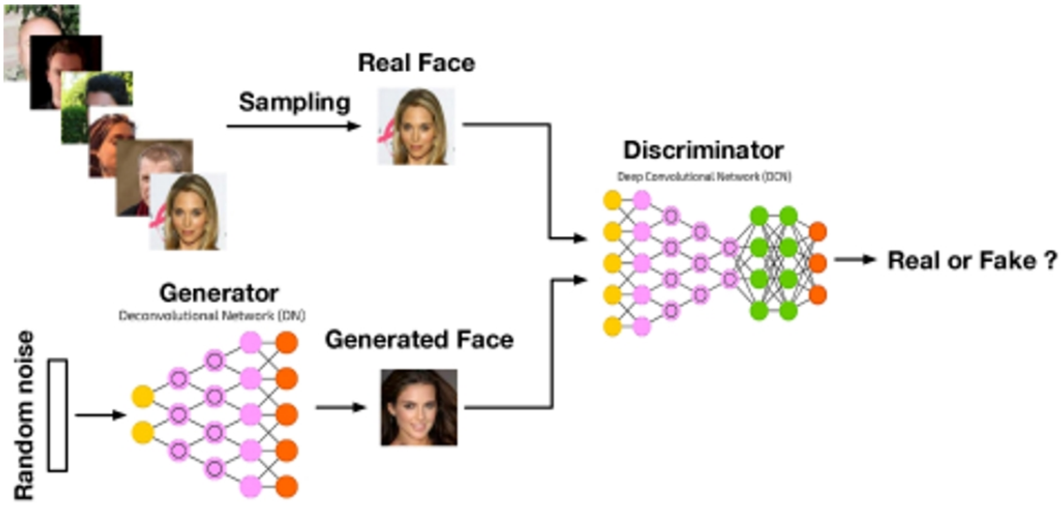


Figure 5.6: Architecture of a GAN. Two deep neural networks (discriminator (D) and generator G)) are synchronously trained during the learning stage. Discriminator is optimized in order to distinguish between real images and generated images while generator is trained for the sake of fooling discriminator from discerning between real images and generated images.

classifier from predicting the correct attribute values. The authors have trained and tested Fader Network considering the CelebA dataset [91], obtaining a model that can significantly change the perceived value of the attributes while preserving the natural appearance of the input images. Tests have also been carried out on the Oxford-102 dataset [107] (containing about 9, 000 images of flowers classified in 102 categories), by changing the colour of the flower while keeping the background unchanged. Excellent results have also been achieved in this test. Fader Network allows also to exchange multiple attributes simultaneously, for example gender, open eyes and glasses attributes. The code is available at `https://github.com/facebookresearch/FaderNetworks`.

In [124], Shen et al. a novel method based on **residual image learning** for face attribute manipulation is proposed. The method combines the generative power of the GAN model with the efficiency of the feed-forward network. It can model the manipulation operation, as learning the residual image, defined as the difference between the original input image and the desired manipulated one. The proposed work focuses on the attribute-specific face area instead of the entire face which contains many redundant and irrelevant details. They develop a dual scheme able to learn two inverse attribute manipulations (one as the primal manipulation and the other as the dual manipulation) simultaneously. For each face attribute manipulation there are two image transformation networks called $G0$ and $G1$ and a discriminative network D. $G0$ and $G1$ that respectively simulate the primal and the dual manipulation. D classifies the reference images and generated images into three categories.

Several DeepFake based techniques that are present at the state of the art are limited regarding the management of more than two domains (for example, to change hair colour, gender, age, and many others features in a face), since they should generate different models for each pair of image domains. A method capable of performing image-to-image translations on multiple domains using a single model has been proposed by Choi et al. [39] by means of a technology called **StarGAN**, a generative adversarial network. The main purpose was to define a scalable image-to-image translation model across multiple domains using a single generator and a discriminator. The authors used two different types of face datasets: CelebA [91] contains 40 labels related to facial attributes such as hair color, gender and age, and RaFD dataset [84] containing 8 labels corresponding to different types of facial expressions ("happy", "sad", etc). Given a random label (for example hair color, facial expression, etc) as input, which represents information about the target domain, the network is able to perform an image-to-image translation operation considering the given label. So, for example, it is possible to carry out a summary of facial expressions of CelebA images using the features learned by RaFD. The results have been compared with other existing methods [87, 111, 154] and show how StarGAN manages to generate images of superior visual quality. The code is available at `https://github.com/yunjey/stargan`.

Wang et al. [137] propose Identity-Preserved Conditional Generative Adversarial Networks (**IPCGAN**), a framework for facial aging. In particular, IPCGAN is composed of three parts: a CGANs module (takes an input image and a target age to generate a new face with that age), an identity-preserved module (guarantees the aged face has the same input identity) and an age classifier (to ensure that the output has the desired age). In this work, the authors consider faces with different ages divided into 5 groups: 11-20, 21-30, 31-40, 41-50 and 50+. Given an image of the face x, they use the C_s information to indicate the age group to which x belongs. The aging of a face aims to generate a synthesized face of the target age group C_t. The framework has been trained and tested using the Cross-Age Celebrity Dataset (CACD) [37]. This dataset contains more than $160,000$ images of the faces of $2,000$ celebrities between the ages of 16 and 62, they all varies in pose, lighting and expression. The performances of IPCGAN have been compared with methods present in the state of the art: age conditional Generative Adversarial Networks (acGAN) [24] and Conditional Adversarial Autoencoder Network (CAAE) [150], methods based on conditional GANs, which achieve performance at vanguard for aging of the face. The qualitative and quantitative tests carried out by the authors show that IPCGAN achieves the best results. For example, the authors highlight the fact that the images generated by acGAN have many artifacts and, in particular, the more the target age grows, the more one risks losing identity. The images generated by CAAE appear blurry and unrealistic. The images obtained through IPCGAN are better: fewer artifacts, higher image quality and less chance of losing identity. Finally, IPCGAN can also be used to perform multi-attribute transfer tasks. The code is available at `https://github.com/dawei6875797/Face-Aging-with-Identity-Preserved-Conditional-Generative-Adversarial-Networks`.

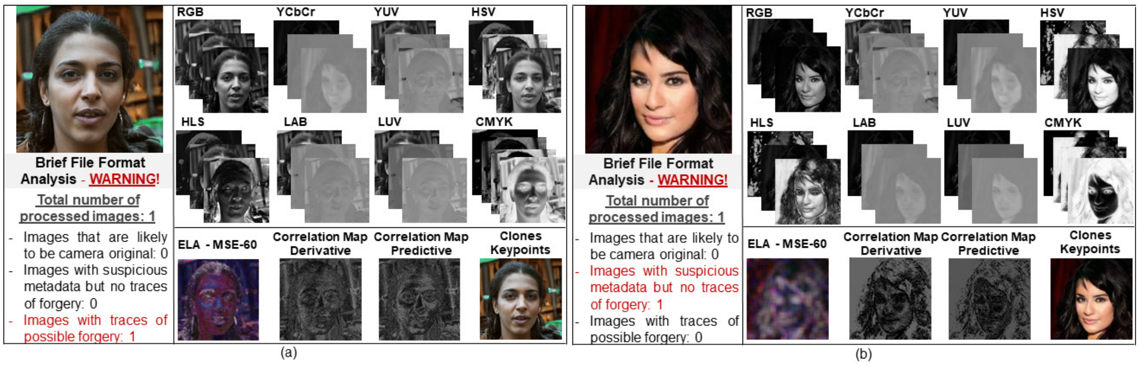
The Style Generative Adversarial Network, or **StyleGAN**, is an extension to the GAN architecture that proposes large changes to the generator model, including the use of a mapping network to map points in latent space to an intermediate latent space, the use of the intermediate latent space to control style at each point in the generator model, and the introduction to noise as a source of variation at each point in the generator model. The resulting model is capable not only of generating impressively photorealistic high-quality photos of faces, but also offers control over the style of the generated image at different levels of detail through varying the style vectors and noise. In December 2018, the visual computing company NVIDIA, released an open source code for photorealistic face generation software created thanks to the StyleGAN algorithm [79]. Then, Uber's computer engineer Phillip Wang created the website `https://thispersondoesnotexist.com/` and then on 11 February 2019 published it on the public group Facebook Artificial Intelligence and Deep Learning. StyleGAN algorithm is to be able to create realistic pseudo-portraits, difficult to judge as fakes. The features and expressions are extremely credible, even if some small details allow us to understand that we are facing a fake. StyleGAN has also difficulty with the definition of the teeth and it cannot identify the backgrounds. Furthermore, there are often fluorescent spots, similar to water drops, which can appear anywhere on the image. The cose is available at `https://github.com/NVlabs/stylegan`.

Recent studies on face attribute transfer have achieved great success. A lot of models are able to transfer face attributes with an input image. In [141] Xiao et al. propose a novel model, **ELEGANT** (the abbreviation of Exchanging Latent Encodings with GAN for Transferring multiple face attributes), which receives two images of opposite attributes as inputs. Their model can transfer exactly the same type of attributes from one image to another by exchanging certain part of their encodings. All the attributes are encoded in a disentangled manner in the latent space, which enables to manipulate several attributes simultaneously. With the help of multi-scale discriminators for adversarial training, it can even generate high-quality images with finer details and less artifacts by comparing with other methods on the CelebA face database. A pytorch implementation is available at `https://github.com/Prinsphield/ELEGANT`.

5.3 Fighting DeepFakes with Image Forensics

Image forgery and alteration is not a new problem introduced by DeepFakes. Counterfeiting an image with image editing tools like Photoshop is still very common as today and the Image/multimedia forensics science as already dealt with the problem [29]. There are several techniques that try to understand if a multimedia content is fake by means of various strategies to detect anomalies in the hidden structure of the multimedia content itself, exploiting noise, compression parameters, etc. Some try to reconstruct the history of the image [64] to identify the source-acquiring device or software, others instead analyze anomalies in the compressed JPEG domain [28, 63, 58] like Galvan et al. [58] who proposed a method which is able to recover the coefficients of the first compression process in a double compressed JPEG image to verify if there are altered elements. Another example from Battiato et al. [28] exploits the statistical distribution of the DCT coefficients in order to detect irregularities due to the presence of a signal overlapped on the original image. More recently, Giudice et al. [63] proposed a new analysis that can be carried out in the DCT domain able to automatically classify doubly compressed JPEG images with extremely high precision, giving forensics experts a tool to find the first evidence of image alteration. Nowadays methods have been created that try to detect anomalies in images and videos using neural networks. Not only the hidden structure of the image can be useful fakeness analysis. Even if the DeepFake images are extremely realistic, the visible contents could be analysed in order to find anomalies useful for detection.

In this context an interesting work is that proposed by Li et al. [88] where they try to detect the lack of blinking of the synthesized faces to define whether the data is a fake by exploiting a deep neural network model that combines CNN and recursive neural network, known as CNN longterm recurrent (LRCN) to distinguish whether the eye is open or closed


Brief File Format Analysis - WARNING!

Total number of processed images: 1

- Images that are likely to be camera original: 0
- Images with suspicious metadata but no traces of forgery: 0
- Images with traces of possible forgery: 1

Brief File Format Analysis - WARNING!

Total number of processed images: 1

- Images that are likely to be camera original: 0
- Images with suspicious metadata but no traces of forgery: 1
- Images with traces of possible forgery: 0

Figure 5.7: Example of Analysis carried out with Amped Authenticate software. (a) Image generated by STYLEGAN. (b) Image generated by STARGAN. In both examples we show only some of the elements analyzed with Amped Authenticate, such as the analyzed different color spaces to understand if anomalies are found; ELA (Error Level Analysis) for Identification of spliced areas of the image that have been compressed differently; Correlation map to analyze and identify the correlation between the pixels of the image; Clones Keypoints to find parts of the image that appear to be cloned.

and consider the temporal structure of the eye beat. An accurate analysis shows that, in general for an adult, the interval between a blink of the eyes and the next one is within $2-10 seconds$, and the typical duration of a blink is $0.1-0.4 seconds/blink^2$. The detection of lack of eye blinking is a possible signal of a DeepFake video. The proposed method was tested with about 50 videos taken from the eye blinking video (EBV) dataset. The network takes as input the patch containing the eyes, this data is obtained with the following operations: 1) Face detection and Face landmark to identify the various regions of the face 2) Distortions removal (they are caused by the movement of the head) 3) Eyes patch extraction (box size: $1.25 horizontal, 1.75 vertical$). Another interesting technique focuses on eye aspect ratio (EAR) [125] produced by GANs that, even if realistic, is too-perfect, asymmetrical or just abnormal.

In this context, Marra et al. [96] discussed the performance of various image-to-image translations detectors, both in ideal conditions and in the presence of compression, performed at the time of uploading to social networks. The study, conducted on a dataset of 36302 images, shows that it is possible to obtain detection accuracy up to 95% both with conventional detectors and with deep learning based detectors, but only the latter continue to provide high accuracy, up to 89%, on compressed data. Many detectors do not work so well with images compressed by social networks such as twitter.

Hsu et al. [69] also proposed an interesting method for detecting DeepFakes, called Deep Forgery Discriminator (DeepFD). Thanks to the implementation of a new discriminator called “contrastive loss” it is possible to find the typical characteristics of the synthesized images generated by different GANs and therefore use a classifier to detect such fake images. For the training phase they used the CelebA dataset [91], considering 5 state-of-the-art GANs to generate the pool of false images: DCGAN (Deep convolutional GAN) [115], WGAP (Wasserstein GAN) [25], WGAN-GP (WGAN with Gradient Penalty) [67], LSGAN (Least Squares GAN) [95] and PGGAN [78]. Using the DeepFD they detected 94.7% of fake images generated by numerous state-of-the-art GANs, exceeding the other basic approaches present in the state of the art in terms of precision and recall rates. The code is available at `https://github.com/jesse1029/Fake-Face-Images-Detection-Tensorflow`.

Next to images, there are some specific techniques for video’s which are try to define if they are DeepFake. A video based DeepFake Detection method is described in the work of Güera et al. [66]. The authors used a CNN to extract frame-level features. In particular they use a Recurrent neural network (RNN) to train a classifier able to recognise if the video was manipulated or not by evaluating temporal inconsistencies introduced between the frames where face is modified. They evaluated the method with 600 videos: 300 fakes found on the web, 300 of real scenes taken from the HOHA dataset [85]. From experiments carried out by the authors, they get 97% accurate for fake detection.

Vishwanath et al [133] propose a method to verify the identity of a person among disguised and impostors images. Their approach is based on VGG-face architecture paired with Contrastive loss based on cosine distance metric. They use the Disguise Faces in Wild (DFW) dataset [80], [48], which is so far the largest dataset of disguised and imposter images, The DFW dataset consists of 1000 identities (400 in training set and 600 in the testing set), with a total of 11155 images. There consists of 3 types of images: genuine images, disguised images and the imposter images. The genuine images are the usual visual image (photograph) of the person, the disguised images is the image of the same identity in disguise and the imposter images are the images consisting of images of other persons who look like the identity. The method exploits, the usage of three different loss functions.

5.3.1 DATASET DEEPFAKE VIDEO

In the state of the art are present different techniques which provide DeepFake video datasets, specifically designed to create new detection methods. Many of these datasets present some problems such as low visual quality and do not resemble the DeepFake videos present on the Internet. For this reason Li et al. [90] created a new DeepFake video dataset called Celeb-DF, containing 590 real videos and 5.639 DeepFake videos. The authors evaluated the performance of some methods in the state of the art of DeepFake detection on video [153, 20, 144, 89, 97, 119, 105, 106] considering the Celeb-DF dataset. The detection performance of these methods degrade with Celeb-DF, despite the excellent results obtained on other datasets. Celeb-DF is available at `https://github.com/danmohaha/celeb-deepfakeforensics`. Finally, Rossler et al. [119] examine different techniques present in the state of the art based on DeepFake. They propose an automated benchmark for fake detection, based mainly on four manipulation methods: two computer graphics-based methods (Face2Face [130], FaceSwap [13]) and 2 learning-based approaches (DeepFakes [12], NeuralTextures [129]). The authors chose these 4 methods mainly to consider different types of data, that is data with random compression and random dimensions in order to obtain the most realistic possible scenario. They addressed the problem of fake detection as a binary classification for each frame of manipulated videos, considering different techniques present in the state of the art (Steg. Features + SVM [57], Cozzolino et al. [42], Bayar and Stamm [30], Rahmouni et al. [116], MesoNet [20], XceptionNet [40]). A face tracking method is initially applied to the input image (Thies et al. [130]) (so as to work only on that particular region) to switch to the classification methods. The experiments conducted show that XceptionNet achieves the best results. They also tested XceptionNet with images without face tracking information. In this case, however, the XceptionNet classifier has significantly lower accuracy. Another major contribution of this work is a new large-scale set of manipulated facial images consisting of over 1.8 million images of $1,000$ videos with real sources. The code is available at `https://github.com/ondyari/FaceForensics`. Fighting DeepFakes with Deep Neural Network techniques have already been demonstrated good performances, both in images and video. But what are the evidences found by those techniques, namely explaining why they classify an image as DeepFake, is still not completely known. All the listed-above studies deal with images and videos created by GANs without any other processing. Marra et al. [96] already showed that JPEG compression could lead to loss of accuracy of detection techniques. Given this, a complete analysis that exploits Deep Neural Network detectors in conjunction with standard (and of today) manual image forensics analysis techniques could lead to a solution of the problem. To this aim, a detailed image forensics analysis will be presented in the next Section.

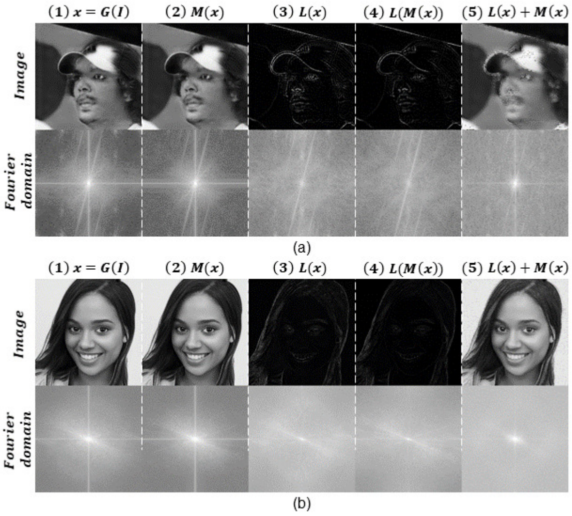


Figure 5.8: Examples of analyzed images generated by (a) STARGAN and (b) STYLEGAN. Each image I of both datasets was converted to grayscale (1) and applied progressively: the Median filter (2), the Laplacian filter (3), the Laplacian filter (4) applied to the result of (2), the sum of the results between the Median and Laplacian filters (5). For each operation performed, we show the Fourier transform.

5.4 DeepFake Forensic Investigation

In the previous paragraphs the most accurate techniques used nowadays to create DeepFake were described, and therefore different detection techniques were discussed. Most of the existing detection techniques are based on neural networks and it is very complicated to explain their results in a courtroom, giving their black-box features. Being able to understand the type of architecture used, the anomalous features and deterministically explaining why an image is a Deepfake is a process that is still almost completely absent in the literature.

In general, in the Multimedia Forensics [29] best practices are used to determine if an image is a fake, where, one tries to understand if it comes from a particular source, analyzing, information present in the metadata (how much present and not altered), the PNRU present in the images (fingerprint of the source), analysis of the coefficients obtained by JPEG compression, analysis in the Fourier domain and much more. If a specific information is not present, the suspicion arises that the data in question is not real. Nowadays, the current neural networks that generate DeepFake do not perform the same operations that are performed by any acquisition imaging source. This leads to obtain images with anomalies at the pixel level, obtaining a pattern that is not present if we consider an image generated by any device or software (camera, scanner, social network, etc.).

A forensics analysis was carried out on sample Deepfake Images by means of one of the most famous image forensics software "Amped Authenticate"[7]: it was employed to check if it is possible to identify whether an image, generated by a GAN, has anomalies, considering in particular the output of two types of technologies: StarGAN [39] and StyleGAN [79], briefly described Section 5.2. In particular we analyzed JPEG structure of the image, we tried to infer Camera Identification (PNRU Identification), and then analysed those images in different color spaces (RGB, YCbCr, YUV, HSV, HLS, XYZ, LAB, LUV, CMYK); domains (ELA, DCT Map, JPEG Dimples Map, Blocking Artifacts, JPEG Ghosts Map, Fusion Map, Correlation Map, PRNU Map, PRNU Tampering, LGA) and by means of many forgery detection techniques (Clones Blocks, Clones Keypoints (Orb and Brisk). Figure 5.7 shows an overview of the results obtained and it is possible to see that in some cases the images may show specific anomalies while in other cases simple warnings are shown by the tool. However, this is not enough to define with certainty if the images are Deepfake, the only thing that could be inferred is that they are probably not-authentic and integrity is broken.

A deep analysis was then carried out in the frequency domain. Indeed, useful information can be obtained by working on Deepfake "candidate" images after the Fourier transform. In general, a simple operation of Forgery or a Deepfake contains "abnormal" frequencies not present in real images. The application of convolutive filters with the respective Fourier

[7] https://ampedsoftware.com/it/authenticate/

spectra highlights the presence of a somewhat suspicious pattern, not present in real data. We performed several tests, using the Laplacian and median filters and the combination of them in order to enhance them (Figure 5.8). In the Fourier domain, as shown in Figure 5.8, it is possible to notice anomalous frequencies that substantially represent a pattern of that particular network used to generate fakes. This information is useful for identifying the type of neural network used and the areas in which that pattern is present. In Figure 5.8 the anomalies are clearly visible and are different for the two samples of each different techniques. Probably these patterns represents the way that deep neural network, and their convolutive layers, create the image, so they could be related to the hyper-parameters such as kernel masks employed in each generative technique. Obviously further investigation is needed but the conjunction of standard image forensics techniques that arises warning of "fakeness" and the detection of known anomalies (different for each technology) in the Fourier domain can achieve good results in terms of Deepfake Detection performance.

There are some methods in the literature that perform frequency domain analysis to detect whether an image is the result of a Deepfake. Zhang et al. [149] propose a method to classify DeepFake images considering the spectra of the frequency domain as input. In general, to be able to classify a data as real or fake, different machine learning methods require a large number of images generated by different GANs in order to train classifiers to distinguish them. Since the artifacts produced during the DeepFake generation pipeline are often common among different GANs (and represented as replicas of spectra in the frequency domain), the authors [149] have proposed a GAN simulator, called AutoGAN, so to emulate the process commonly shared by popular GAN models in the generation of DeepFakes. From the results obtained by the authors, considering images simulated via AutoGAN to train a classifier based on spectra in the frequency domain, this approach achieves very good performance for what regards the detection of fake images generated by popular GAN models. The code is available at `https://github.com/ColumbiaDVMM/AutoGAN`. Durall et al. [49] present a method for DeepFake detection based on the analysis in the frequency domain. To evaluate the technique, the authors combined high-resolution face images taken from different public datasets of real faces (CelebA-HQ data set [78], Flickr-Faces-HQ data set [79]) and fakes (100K Faces project [11], this person does not exist [14]), creating a new dataset called Faces-HQ. From the carried out classification tests, the authors achieved an accuracy of 100% using supervised classifiers (SVM, Logistic Regression) and of 96% using an unsupervised classifier (K-means). The code is available at `https://github.com/cc-hpc-itwm/DeepFakeDetection`.

The application of convolutive filters with the respective Fourier spectra highlights the presence of a somewhat suspicious pattern, not present in real data. We performed several tests, using the Laplacian and median filters and the combination of them in order to enhance them Figure 5.8 shows an example.

In the Fourier domain, as shown in Figure 5.8, it is possible to notice anomalous frequencies that substantially represent a pattern of that particular network used to generate fakes. This information is useful for identifying the type of neural network used and the areas in which that pattern is present. In Figure 5.8 the anomalies are clearly visible and are different for the two samples of each different techniques. Probably these patterns represents the way that deep neural network, and their convolutive layers, create the image, so they could be related to the hyper-parameters such as kernel masks employed in each generative technique. Obviously further investigation is needed but the conjunction of standard image forensics techniques that arises warning of "fakeness" and the detection of known anomalies (different for each technology) in the Fourier domain can achieve good results in terms of

DeepFake Detection performance.

Forensics vs. anti-forensics is always an open game, and while we are dealing with detection methods, there is already a first attempt that seeks to hide those anomalies that were described above by introducing information, such as the digital fingerprint of the camera, into the fake images. This method was already proposed by Cozzolino et al. [43], in his work called "SpoC: Spoofing Camera Fingerprints". The SpoC authors propose a GAN-based approach capable of injecting traces of the camera into synthetic images (thus removing traces of the synthesis process), tricking avant-garde detectors into believing that the image was acquired from that model, achieving good results in a wide range of conditions.

Chapter 6

Conclusions

In this thesis we deal with rich ecosystem, that will be further investigate from the scientific community in next years. This chapter summarizes activities and results achieved. It describes the development of our works according to the three issues addressed during the Ph.D course by highlighting some contexts more and more central in our everyday life. It mainly concerns the Social and Multimedia Forensics field,passing through a detection of false multimedia content and manipulated images (focusing particularly on the double compressed JPEG images), until arriving to the emerging phenomenon of the DeepFakes. For each topic it carries out an analysis of the state of the art, analyzes possible approaches for solving problems in such research areas and proposes possible solutions.

In the introductory part it mentions the usability of digital evidences at Court in the judicial proceedings and lists the main international standards in Digital Forensic Domain.

The second chapter presents an overview of social network platform and forensic analysis about defamation, analyzes how investigate on social media crime and it shows the results over the 250 different existing social.

In the next two ones we focus attention on JPEG Compression and Forensics on Double Compressed JPEG Images with application of Galvan's Algorithm in Different Scenarios. Unfortunately in this field results are not consistent but many point remain open for future work and discussion.

The last chapter describes the critic and recent phenomenon of Deepfake. It presents a brief overview of technologies able to produce Deepfake images of faces, a forensics analysis of those images with standard methods and a preliminary idea on how to fight Deepfake images of faces by analysing anomalies in the frequency domain.

The prevalence of digital devices has caused digital multimedia content to become pervasive throughout modern society. However, because digital content can be easily altered using widely available software, its authenticity must be established before it can be trusted. As a result, a number of digital forensic techniques have been developed but Forensics vs. anti-forensics is always an open game, and while we are dealing with detection methods, there is already a first attempt that seeks to hide those anomalies.

Part of this thesis work has been published in the following papers:

C. Nastasi, L. Guarnera, O. Giudice, S. Battiato. Preliminary Forensics Analysis of Deepfake Images. Conference Proceedings of AEIT 2020.

L. Guarnera, O. Giudice, C. Nastasi, S. Battiato. Forensics Analysis of DeepFake Images. Arxiv Preprint, ArXiv 2004-12626.

S. Battiato, O. Giudice, L. Guarnera, C. Nastasi. Analisi Forense di immagini Deepfake di volti. Chapter on IISFA Memberbook 2020-2021.

C. Nastasi, S. Battiato. Defamation 2.0: new Threats in Digital Media Era. An Overview on Forensics Approaches in the Social Network Ecosystem. Submitted for publication on Conference Proceedings of IMPROVE 2021.

Bibliography

[1] https://socialmedialist.org/elenco-dei-social-network-nel-mondo.html.

[2] Electronic CSI, A Guide for First Responders, 2nd edition, National Institute of Justice, April 2008.

[3] https://www.iso.org/obp/ui/es/#iso:std:iso-iec:27037:ed-1:v1:en.

[4] https://www.iso.org/obp/ui/#iso:std:iso-iec:27042:ed-1:v1:en.

[5] https://www.lastampa.it/cultura/2018/02/03/news/quanti-social-network-esistono-1.33975738.

[6] https://en.wikipedia.org/wiki/List_of_social_networking_websites.

[7] https://socialmedialist.org/social-media-apps.html.

[8] https://en.wikipedia.org/wiki/JPEG.

[9] https://en.wikipedia.org/wiki/Joint_Photographic_Experts_Group.

[10] Dataset Eastman Kodak Company: PhotoCD PCD0992,[Online]: http://r0k.us/graphics/kodak/.

[11] 100,000 faces generated. `https://generated.photos/`. Accessed: 2020-02-26.

[12] Deepfakes. `https://github.com/deepfakes/faceswap/`. Accessed: 2019-11-27.

[13] Faceswap. `https://github.com/MarekKowalski/FaceSwap/`. Accessed: 2019-11-27.

[14] This person does not exist. `https://thispersondoesnotexist.com/`. Accessed: 2020-02-26.

[15] Warren weaver. `https://en.wikipedia.org/wiki/Warren_Weaver/`. Accessed: 2019-12-4.

[16] G. Desjardins H. Soyer J. Kirkpatrick-K. Kavukcuoglu R. Pascanu R. Hadsell A. A. Rusu, N. C. Rabinowitz. Progressive neural networks. 2016.

[17] K. Simonyan A. Brock, J. Donahue. Large scale gan training for high fidelity natural image synthesis. 2018.

[18] G. Hinton A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[19] L. Metz A. Radford and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. page 2015.

[20] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[21] Noora Al Mutawa, Ibrahim Baggili, and Andrew Marrington. Forensic analysis of social networking applications on mobile devices. *Digital Investigation*, 9:S24 – S33, 2012. The Proceedings of the Twelfth Annual DFRWS Conference.

[22] I. Mohammed Al-Saleh and Yahya A. Forihat. Skype forensics in android devices. In *International Journal of Computer Applications*, volume 78, pages 38–44, September 2013.

[23] Sameera Almulla, Youssef Iraqi, and Andrew Jones. A state-of-the-art review of cloud forensics. In *Journal of Digital Forensics, Security and Law(JDFSL)*, volume 9, May 2014.

[24] G. Antipov, M. Baccouche, and J. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[25] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[26] M. Baca and Z. Cosic, J.and Cosic. Forensic analysis of social networks (case study). In *Information Technology Interfaces (ITI), Proceedings of the ITI 2013 35th International Conference on*, pages 219–223, 2013.

[27] M. Barni, A. Costanzo, and Sabatini L.:. Identification of cut & paste tampering by means of double-JPEG detection and image segmentation. In *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS*, pages 1687–1690, 2010.

[28] S. Battiato and Messina G.:. Digital forgery estimation into DCT domain: a critical analysis. In *Proceedings of the First ACM workshop on Multimedia in forensics*, pages 389–399, 6, 2010. 25.

[29] Sebastiano Battiato, Oliver Giudice, and Antonino Paratore. Multimedia forensics: discovering the history of multimedia contents. In *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*, pages 5–16. ACM, 2016.

[30] B. Bayar and M. C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016.

[31] T. Bianchi and Piva A.:. Detection of non-aligned double JPEG compression with estimation of primary compression parameters. In *18th IEEE International Conference on Image Processing (ICIP*, pages 1929–1932, 2011.

[32] T. Bianchi and Piva A. Image forgery localization via block-grained analysis of JPEG artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, 2012.

[33] T. Bianchi and A. de Rosa. and piva a.: Improved DCT coefficient analysis for forgery localization in JPEG images. In Ieee International, editor, *Conference on Acoustics Speech and Signal Processing*, 2444, 2447, 2011. ICASSP).

[34] danah m. boyd and Nicole B. Ellison. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 10 2007.

[35] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[36] F. Huszar J. Caballero A. Cunningham A. Acosta-A. Aitken A. Tejani J. Totz Z. Wang et al. C. Ledig, L. Theis. Photorealistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–114, 2017.

[37] B. Chen, C. Chen, and W. H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer, 2014.

[38] David Chiang. A hierarchical phrase-based model for statistical machine translation.

[39] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[40] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[41] Ferenc Huszar Jose Caballero Andrew Cunningham Alejandro Acosta-Andrew Aitken Alykhan Tejani Johannes Totz Zehan Wang Wenzhe Shi Christian Ledig, Lucas Theis. Photo-realistic single image super-resolution using a generative adversarial network. 2014.

[42] D. Cozzolino, G. Poggi, and L. Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164, 2017.

[43] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva. Spoc: Spoofing Camera Fingerprints. *arXiv preprint arXiv:1911.12069*, 2019.

[44] Nandita Dalmia and Manish Okade. A novel technique for misalignment parameter estimation in double compressed jpeg images. In *Visual Communications and Image-Processing, VCIP*, pages 1–4, 11 2016.

[45] Nandita Dalmia and Manish Okade. First quantization matrix estimation for double compressed jpeg images utilizing novel dct histogram selection strategy. In *ICVGIP '16 Proceedings of the Tenth Indian Conference on Computer Vision*. Graphics and Image Processing Article No. 27, 2017.

[46] Nandita Dalmia and Manish Okade. Robust first quantization matrix estimation based on filtering of recompression artifacts for non-aligned double compressed jpeg images. *Signal Processing: Image Communication*, 61:9–20, 2018.

[47] E. L Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015.

[48] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PloS one*, 9(7), 2014.

[49] R. Durall, M. Keuper, F. Pfreundt, and J. Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

[50] Kyunghyun Cho Dzmitry Bahdanau and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2014.

[51] A. Szlam E. Denton, S. Chintala and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

[52] R. Fergus al E. L. Denton, S. Chintala. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, pages 1486–1494, 2015.

[53] Phillip Isola Jun-Yan Zhu Tinghui Zhou Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Berkeley AI Research (BAIR) Laboratory, UC Berkeley*, 2018.

[54] Z. Fan and de Queiroz Ricardo L.:. Maximum likelihood estimation of JPEG quantization table in the identification of bitmap compression history. In *Proceedings of the International Conference on Image Processing*, pages 948–951. 1, 2000.

[55] Z. Fan and de Queiroz Ricardo L. Identification of bitmap compression history: JPEG detection and quantizer estimation. *IEEE Transactions on Image Processing*, 12(2):230–235, 2003.

[56] Wei Lu Hongmei Liu Fei Xue, Ziyi Ye and Bin Li. Mse period based estimation of first quantization step in double compressed jpeg images. *Signal Processing: Image Communication*, 57:76–83, 2017.

[57] J. Fridrich and J. Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012.

[58] F. Galvan, G. Puglisi, A. R. Bruna, and Battiato S.:. First quantization coefficient extraction from double compressed JPEG images. In *International Conference on Image Analysis and Processing (ICIAP*, pages 783–792, 2013.

[59] F. Galvan, G. Puglisi, A. R. Bruna, and Battiato S. First quantization matrix estimation from double compressed JPEG images. *IEEE Transactions on Information Forensics and Security*, 9(8):1299–1310, 2014.

[60] Fausto Galvan, Giovanni Puglisi, Arcangelo Ranieri Bruna, and Sebastiano Battiato. First quantization matrix estimation from double compressed JPEG images. *IEEE Transactions on Information Forensics and Security*, 9(8):1299–1310, 2014.

[61] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition*, Winter semester, 2014.

[62] O. Giudice, Paratore A.B., Moltisanti, and S. Battiato. A classification engine for image ballistics of social data. In *International Conference on Image Analysis and Processing*, pages 625–636, 2017.

[63] O. Giudice, F. Guarnera, A. Paratore, and S. Battiato. 1-D DCT domain analysis for JPEG double compression detection. In *Proceeedings of International Conference on Image Analysis and Processing*, pages 716–726. Springer, 2019.

[64] O. Giudice, A. Paratore, M. Moltisanti, and S. Battiato. A classification engine for image ballistics of social data. In *Proceeedings of International Conference on Image Analysis and Processing*, pages 625–636. Springer, 2017.

[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[66] D. Güera and E. J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[67] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[68] H. Li S. Zhang X. Wang X. Huang-D. N. Metaxas H. Zhang, T. Xu. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, page 5907–5915, 2017.

[69] C. Hsu, C. Lee, and Y. Zhuang. Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 388–391. IEEE, 2018.

[70] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian J. Goodfellow, Jean Pouget-Abadie. Generative adversarial nets. *NIPS*, 2014.

[71] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[72] Quoc V. Le Ilya Sutskever, Oriol Vinyals. Sequence to sequence learning with neural networks. In Canada Montreal, editor, *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112. Volume 2, December 08-13 2014.

[73] P. Isola J.-Y. Zhu, T. Park and A. A. Efros. Unpaired image-toimage translation using cycle-consistent adversarial networks. 2017.

[74] M. Mathieu J. Zhao and Y. LeCun. Energy-based generative adversarial network. 2016.

[75] Jebara. Discriminative learning: ”this distinction between conditional learning and discriminative learning is not currently a well established convention in the field.”. 2004.

[76] Tian Junjing, Bi Yan, and Ma Jinqiang. Research on forensics of social network relationship based on big data. In *Journal of Physics: Conference Series 1584 012022. DMCIT 2020*, 2020.

[77] B. B. Meshram K. K. Sindhu. Digital forensic investigation tools and procedures. In *International Journal of Computer Network and Information Security(IJCNIS)*, volume 4, pages 39 – 48, May 2012.

[78] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[79] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[80] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2018.

[81] Caglar Gulcehre Dzmitry Bahdanau Fethi Bougares Holger Schwenk Yoshua Bengio KyungHyun Cho, Bart van Merrienboer. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar*, pages 1724–1734, October 25-29 2014.

[82] Dzmitry Bahdanau KyungHyun Cho, Bart van Merrienboer and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. 2014.

[83] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[84] O. Langner, R. Dotsch, G. Bijlstra, D. HJ Wigboldus, S. T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[85] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[86] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

[87] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[88] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[89] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[90] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019.

[91] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[92] R. Fergus M. D. Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[93] A. Mahajan, M. S. Dahiya, and H. P. Sanghvi. Forensic analysis of instant messenger applications on android devices. In *arXiv preprint arXiv:1304.4915*, 2013.

[94] Asma Majeed, Haleemah Zia, Rabeea Imran, and Shahzad Saleem. Forensic analysis of three social media apps in windows 10. In *2015 12th International Conference on*

High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), pages 75–79, 2015.

[95] X. Mao, Q. Li, H. Xie, R. YK Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[96] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018.

[97] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[98] Josè A. R. Fonollosa Maxim Khalilov. Syntax-based reordering for statistical machine translation. *Computer Speech and Language*, 4(25):761–788, 2011.

[99] M. Mirza and S. Osindero. Conditional generative adversarial nets. 2014.

[100] S. Mohtasebi and A. Dehghantanha. Defusing the hazards of social network services. In *Int. J. Digit. Inf. Wirel. Commun.*, pages 504–516, 2011.

[101] M. Moltisanti, A. Paratore, S. Battiato, and L. Saravo. Image manipulation on facebook for forensics evidence. In *International Conference on Image Analysis and Processing*, pages 506–517, 2015.

[102] Phil Blunsom Nal Kalchbrenner. Recurrent continuous translation models. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1700–1709. Association for Computational Linguistics, 2013.

[103] Jordan Ng. Discriminative classifiers model the posterior p(y—x) directly, or learn a direct map from inputs x to the class labels. 2002.

[104] Jordan Ng. Generative classifiers learn a model of the joint probability p(x,y)of the inputs x and the label y, and make their predictions by using bayes rules to calculate p(y—x) and then picking the most likely label y. 2002.

[105] H. H Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.

[106] H. H Nguyen, J. Yamagishi, and I. Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019.

[107] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[108] T. Brox O. Ronneberger, P. Fischer. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[109] Franz Josef Och. Minimum error rate training for statistical machine translation.

[110] E. Adelson P. Burt. The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540, 1983.

[111] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M Álvarez. Invertible conditional GANs for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[112] Stephen A Della Pietra Peter F Brown, Vincent J Della Pietra and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[113] Franz Josef Och Philipp Koehn and Daniel Marcu. Statistical phrase-based translation. *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, I:48–54, 2003.

[114] Alessio Plebe and Giorgio Grasso. The unbearable shallow understanding of deep learning. *Minds and Machines*, 29(4):515–553, 2019.

[115] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[116] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017.

[117] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[118] Anderson Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne R. B. Carvalho, and Efstathios Stamatatos. Authorship attribution for social media forensics. In *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, 2016.

[119] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019.

[120] A. A Rusu, Neil C R., G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[121] E. Simo-Serra S. Iizuka and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4), 2017.

[122] X. Yan L. Logeswaran B. Schiele H. Lee S. Reed, Z. Akata. Generative adversarial text to image synthesis. 2016.

[123] G. Schaefer and Stich M.: Ucid:. an uncompressed color image database, electronic imaging. *International Society for Optics and Photonics*, pages 472–480, 2003.

[124] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4030–4038, 2017.

[125] T. Soukupová and J. Cech. Eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia*, 2016.

[126] S. Laine T. Karras, T. Aila and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2017.

[127] N. S. Thakur. Forensic analysis of whatsapp on android smartphones. In *University of New Orleans Theses and Dissertations*, 2013.

[128] Florent Retraint Thi-Ngoc-Canh Doan Thanh Thai, Remi Cogranne. Jpeg quantization step estimation and its applications to digital image forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):123–133, 2017.

[129] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[130] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[131] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.

[132] L. Verdoliva. Media forensics and deepfakes: an overview. *arXiv preprint arXiv:2001.06564*, 2020.

[133] S. Vishwanath Peri and A. Dhall. Disguisenet: a contrastive approach for disguised face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–31, 2018.

[134] Luo W.:. JPEG error analysis and its applications to digital image forensics. *IEEE Transactions on Information Forensics and Security*, 5(3):480–491, 2010.

[135] Ibrahim; Marrington Andrew; Moore-Jason; Breitinger Frank Walnycky, Daniel; Baggili. Network and device forensic analysis of android social-messaging applications. In *DIGITAL INVESTIGATION Impact Factor: 0.99*, 2015.

[136] W. Wang, J. Dong, and Tan T.:. Exploring DCT coefficient quantization effects for local tampering detection. *IEEE Transactions on Information Forensics and Security*, 9(10):1653–1666, 2014.

[137] Z. Wang, X. Tang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7939–7947, 2018.

[138] Zhengwei Wang, Qi She, and T. Ward. Generative adversarial networks: A survey and taxonomy. *ArXiv*, abs/1906.01529, 2019.

[139] David Cox William Lotter, Gabriel Kreiman. Unsupervised learning of visual structure using predictive generative networks. 2014.

[140] Schuster Mike Chen Zhifeng Le Quoc Macherey Wolfgang Krikun Maxim Cao Yuan Gao Qin Macherey Klaus Klingner Jeff Shah Apurva Johnson Melvin Liu Xiaobing Kaiser ukasz Gouws Stephan Kato Yoshikiyo Kudo Taku Kazawa Hideto Dean Jeffrey Wu, Yonghui. Google's neural machine translation system: Bridging the gap between human and machine translation. 2016.

[141] T. Xiao, J. Hong, and J. Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 168–184, 2018.

[142] Chen Y.-l. and Hsu C.-t.:. Detecting recompression of JPEG images via periodicity analysis of compression artifacts for tampering detection. *IEEE Transactions on Information Forensics and Security*, 6(2):396–406, 2011.

[143] Y. Bengio P. Haffner al. Y. LeCun, L. Bottou. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324. vol. 86, no. 11, 1998.

[144] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[145] Mohd Najwadi Yusoff, Ali Dehghantanha, and Ramlan Mahmod. Forensic investigation of social media and instant messaging services in firefox os: Facebook, twitter, google+, telegram, openwapp and line as case studies. In *Contemporary Digital Forensic Investigations Of Cloud And Mobile*, pages 41–62, 2017.

[146] Lin Z.:. *First JPEG Quantization Matrix Estimation Based on Histogram Analysis.* Pattern Recognition, 2013.

[147] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[148] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[149] X. Zhang, S. Karaman, and S. Chang. Detecting and simulating artifacts in GAN fake images. *arXiv preprint arXiv:1907.06515*, 2019.

[150] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.

[151] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[152] Feng Wang Bo Xu Zhen Yang, Wei Chen. Improving neural machine translation with conditional sequence generative adversarial nets. In *Proceedings of NAACL-HLT 2018, New Orleans, Louisiana*, pages 1346–1355, June 1 - 6 2018.

[153] P. Zhou, X. Han, V. I Morariu, and L. S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.

[154] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

List of Figures